Evaluating Distant Supervision for Subjectivity and Sentiment Analysis on Arabic Twitter Feeds

Eshrag Refaee and Verena Rieser
Interaction Lab, School of Mathematical and Computer Sciences,
Heriot-Watt University,
EH14 4AS Edinburgh, United Kingdom.
eaar1@hw.ac.uk, v.t.rieser@hw.ac.uk

Abstract

Supervised machine learning methods for automatic subjectivity and sentiment analysis (SSA) are problematic when applied to social media, such as Twitter, since they do not generalise well to unseen topics. A possible remedy of this problem is to apply distant supervision (DS) approaches, which learn from large amounts of automatically annotated data. This research empirically evaluates the performance of DS approaches for SSA on Arabic Twitter feeds. Results for emoticon- and lexicon-based DS show a significant performance gain over a fully supervised baseline, especially for detecting subjectivity, where we achieve 95.19% accuracy, which is a 48.47% absolute improvement over previous fully supervised results.

1 Introduction

Subjectivity and sentiment analysis (SSA) aims to determine the attitude of an author with respect to some topic, e.g. objective or subjective, or the overall contextual polarity of an utterance, e.g. positive or negative. Previous work on automatic SSA has used manually annotated gold standard data sets to analyse which feature sets and models perform best for this task, e.g. (Wilson et al., 2009; Wiebe et al., 1999). Most of this work is in English, but there have been first attempts to apply similar techniques to Arabic, e.g. (Abdul-Mageed et al., 2011; Mourad and Darwish, 2013). While these models work well when tested using cross-validation on limited static data sets, our previous results reveal that these models do not generalise to new data sets, e.g. collected at a later point in time, due to their limited coverage (Refaee and Rieser, 2014). While there is a growing interest within the NLP community in building Arabic corpora by harvesting the web, e.g. (Al-Sabbagh and Girju, 2012; Abdul-Mageed and Diab, 2012; Zaidan and Callison-Burch, 2013), these resources have not been publicly released yet and only small amounts of these data-sets are (manually) annotated. We therefore turn to an approach known as *distant supervision* (DS), as first proposed by (Read, 2005), which uses readily available features, such as emoticons, as noisy labels in order to efficiently annotate large amounts of data for learning domain-independent models. This approach has been shown to be successful for English SSA, e.g. (Go et al., 2009), and SSA for under-resourced languages, such as Chinese (Yuan and Purver, 2012).

The contributions of this paper are as follows: we first collect two large corpora using emoticons and lexicon-based features as noisy labels, which we plan to release as part of this submission. Second, this work is the first to apply and empirically evaluate DS approaches on Arabic Twitter feeds. We find that DS significantly outperforms fully supervised SSA on our held-out test set. However, compared to a majority baseline, predicting negative sentiment proves to be difficult using DS approaches. Third, we conduct an error analysis to critically evaluate the results and give recommendations for future directions.

2 Arabic Twitter SSA Corpora

We start by collecting three corpora at different times over one year to account for the cyclic effects of topic change in social media (Eisenstein, 2013). Table 1 shows the distributions of labels in our data-sets:

1. A gold standard data-set which we use for training and evaluation (spring 2013);
2. A data-set for DS using emoticon-based queries (autumn 2013);
3. Another data-set for DS using a lexicon-based approach (winter 2014).

Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP), pages 174–179,
October 25, 2014, Doha, Qatar.

Data set	Neutral	Polar	Positive	Negative	Total
Gold standard training	1,157	937	470	467	3,031
Emoticon-based training	55,076	62,466	32,842	33,629	184,013
Lexicon-based training	55,076	55,538	18,442	5,013	134,069
Manually labelled test	422	579	278	301	1,580

Table 1: Sentiment label distribution of the gold standard manually annotated and distant supervision data sets.

Gold Standard Data-set: We harvest two gold standard data sets at different time steps, which we label manually. We first harvest a data set of 3,031 multi-dialectal Arabic tweets randomly retrieved over the period from February to March 2013. We use this set as a training set for our fully supervised approach. We also manually label 1,580 tweets collected in autumn 2013, which we use as an independent held-out test set. Two native speakers were recruited to manually annotate the collected data for subjectivity and sentiment, where we define sentiment as a positive or negative emotion, opinion or attitude, following (Wilson et al., 2009). Our gold standard annotations reached a weighted $\kappa = 0.76$, which indicates reliable annotations (Carletta, 1996). We also automatically annotate the corpus with a rich set of linguistically motivated features using freely available processing tools for Arabic, such as MADA (Nizar Habash and Roth, 2009), see Table 2. For more details on gold standard corpus annotation please see (Refaee and Rieser, 2014). [1]

Type	Feature-sets
Morphological	diacritic, aspect, gender, mood, person, part_of_speech (POS), state, voice, has_morphological_analysis.
Syntactic	n-grams of words and POS, lemmas, including bag_of_words (BOW), bag_of_lemmas.
Semantic	has_positive_lexicon, has_negative_lexicon, has_neutral_lexicon, has_negator, has_positive_emoticon, has_negative_emoticon.

Table 2: Annotated Feature-sets

Emoticon-Based Queries: In order to investigate DS approaches to SSA, we also collect a much larger data set of Arabic tweets, where we use emoticons as noisy labels, following e.g. (Read, 2005; Go et al., 2009; Pak and Paroubek, 2010; Yuan and Purver, 2012; Suttles and Ide, 2013). We query Twitter API for tweets with variations of positive and negative emoticons to obtain pairs of micro-blog texts (statuses) and using emoticons as author-provided emotion labels. In following (Purver and Battersby, 2012; Zhang et al., 2011; Suttles and Ide, 2013), we also utilise some sentiment-bearing hash tags to query emotional tweets, e.g. فرح *happiness* and حزن *sadness*. Note that emoticons and hash-tags are merely used to collect and build the training set and were replaced by the standard (positive/ negative) labels. In order to collect neutral instances, we query a set of official news accounts, following an approach by (Pak and Paroubek, 2010). Examples of the accounts queried are: BBC-Arabic, Al-Jazeera Arabic, SkyNews Arabia, Reuters Arabic, France24-Arabic, and DW Arabic. We then automatically extract the same set of linguistically motivated features as for the gold standard corpus.

Emoticon	**Sentiment label**
:) , :-) , :)), (: , (-: , ((:	positive
:(, :-(, :((, :((,): ,)):)-:	negative

Table 3: Emoticons used to automatically label the training data-set.

Lexicon-Based Annotation: We also investigate an alternative approach to DS, which combines rule-driven lexicon-based SSA, e.g. (Taboada et al., 2011), with machine learning approaches, following (Zhang et al., 2011). We build a new training dataset by combining three lexica. We first exploit two existing subjectivity lexica: a manually annotated Arabic subjectivity lexicon (Abdul-Mageed and Diab, 2012) and a publicly available English subjectivity lexicon, MPQA (Wilson et al., 2009), which we automatically translate using Google Translate, following a

[1]This GS data-set has been shared via a special LREC repository available at http://www.resourcebook.eu/shareyourlr/index.php

similar technique to (Mourad and Darwish, 2013). The translated lexicon is manually corrected by removing translations with neutral or no clear sentiment indicator.[2] 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 (duplicates removed). In order to obtain automatic labels for positive and negative instances, we follow a simplified version of the rule-based aggregation approach of Taboada et al. (2011). First, all lexicons and tweets are lemmatised. For each tweet, matched sentiment words are marked with either (+1) or (-1) to incorporate the semantic orientation of individual constituents. This achieves a coverage level of 76.62% (which is computed as a percentage of tweets with at least one lexicon word) using the combined lexicon. The identified sentiment words are replaced by place-holders to avoid bias. To account for negation, we reverse the polarity (switch negation) following (Taboada et al., 2011). The sentiment orientation of the entire tweet is then computed by summing up the sentiment scores of all sentiment words in a given tweet into a single score that automatically determines the label as being: positive or negative. Instances where the score equals zero are excluded from the training set as they represent mixed-sentiment instances with an even number of sentiment words. We validate this lexicon-based labelling approach against a separate development set by comparing the automatically computed labels against manually annotated ones, reaching an accuracy of 69.06%.

3 Classification Experiments Using Distant Supervision

We experiment with a number of machine learning methods and we report the results of the best performing scheme, namely Support Vector Machines (SVMs), where we use the implementation provided by WEKA (Witten and Frank, 2005). We report the results on two metrics: F-score and accuracy. We use paired t-tests to establish significant differences ($p < .05$). We experiment with different feature sets and report on the best results (*Bag-of-Words (BOW)* + *morphological* + *semantic*). We compare our results against a majority baseline and against a fully supervised approach. It is important to mention the most prominent previous work on SSA of Arabic tweets like (Abdul-Mageed et al., 2012) who trained SVM classifiers on a nearly 3K manually labelled data-set to curry out two-stage binary classification attaining accuracy up to 65.87% for the sentiment classification task. In a later work, (Mourad and Darwish, 2013) employ SVM and Naive Bayes classifiers trained on a set of 2,300 manually labelled Arabic tweets. With 10-fold cross-validation settings, the author reported an accuracy score of 72.5% for the sentiment classification task (positive vs. negative).

We evaluate the approaches on a separate held-out test-set that is collected at a later point in time, as described in Section 2.

3.1 Emoticon-Based Distant Supervision

We first evaluate the potential of exploiting training data that is automatically labelled using emoticons. The results are summarised in Table 4.

Polar vs. neutral: The results show a significant improvement over the majority baseline, as well as over the classifier trained on the gold standard data set: We achieve 95.19% accuracy on the held-out set, which is a 48.47% absolute improvement over our previous fully supervised results. We attribute this improvement to two factors. First, the emoticon-based data set is about 60 times bigger than the gold standard data set (see Table 1) and thus the emoticon-based model better generalises to unseen events. Note that this performance is comparable with (Suttles and Ide, 2013) who achieved up to 98% accuracy using emoticon-based DS on English tweets using 5.9 million tweets. Second, neutral instances were sampled from news accounts, which are mainly written in modern standard Arabic (MSA), whereas we assume that tweets including emoticons (which we use for acquiring polar instances) are mainly written in dialectal Arabic (DA). In future work, we plan to investigate this hypothesis further by automatically detecting MSA/DA for a given tweet, e.g. (Zaidan and Callison-Burch, 2013). Abdul-Mageed et al. (2012) show that having such a feature can result in no significant impact on the overall performance of both subjectivity and sentiment analysis tasks.

Positive vs. negative: For sentiment classification, the performance of the emoticon-based approach degrades notably to 51%, which is still

[2] For instance, *the day of judgement* is assigned with a negative label while its Arabic translation is neutral considering the context-independent polarity.

Data-set	majority baseline		fully supervised		emoticon DS		lexicon-presence		lexicon-aggr.	
	F	Acc.	F	Acc.	F	Acc.	F	Acc.	F	Acc.
polar vs. neutral	0.69	53.0	0.43	46.62	**0.95**	**95.19**	**0.95**	**95.09**	0.91	91.09
positive vs. negative	**0.67**	50.89	0.41	49.65	0.51	51.25	0.53	**57.06**	0.52	52.98

Table 4: 2-level and single-level SSA classification using distant supervision (DS).

significantly better that the fully supervised baseline, but nevertheless worse than a simple majority baseline. These results are much lower than previous results on emoticon-based sentiment analysis on English tweets by (Go et al., 2009; Bifet and Frank, 2010) which both achieved around 83% accuracy. The confusion matrix shows that mostly negative instances are misclassified as positive, with a very low recall on negative instances, see Table 5. Next, we investigate possible reasons in a detailed error analysis.

Data set	Precision	Recall
emoticon DS		
positive	0.479	0.81
negative	0.556	0.212
lexicon-presence DS		
positive	0.521	0.866
negative	0.733	0.317
lexicon-aggregation DS		
positive	0.496	0.650
negative	0.583	0.426

Table 5: Recall and precision for Sentiment Analysis

3.1.1 Error Analysis for Emoticon-Based DS

In particular, we investigate the use of sarcasm and the direction emoticons face in right-to-left alphabets.

Use of sarcasm and irony: Using an emoticon as a label is naturally noisy, since we cannot know for sure the intended meaning the author wishes to express. This is especially problematic when emoticons are used in a sarcastic way, i.e. their intended meaning is the opposite of the expressed emotion. An example from our data set is:

(1) جميل يا اهلي :(*great job Ahli :(— referring to a famous football team.*

Research in psychology shows that up to 31% of the time, emoticons are used sarcastically (Wolf, 2000). In order to investigate this hypothesis we manually labelled a random sample of 303 misclassified instances for neutral, positive, negative, as well as sarcastic, mixed and unclear sentiments, see Table 6. Interestingly, the sarcastic instances represent only 4.29%, while tweets with mixed (positive and negative) sentiments represent 5.94% of the manually annotated sub-set. In 34.32% of the instances, the manual labels have matched the automatic emoticon-based labels. Surprisingly, automatic emoticon-based label contrasts the manual labels in 36.63% of the instances. Instances labelled as neutral represent 4.95%. The rest of the instances were assigned ‘unclear sentiment orientation’.

Emoticon Label	Predicted label	Manual label	# instances
Positive	Negative	Mixed	8
Negative	Positive	Mixed	10
Positive	Negative	Negative	59
Negative	Positive	Negative	42
Positive	Negative	Neutral	29
Negative	Positive	Neutral	7
Positive	Negative	Positive	62
Negative	Positive	Positive	52
Positive	Negative	Sarcastic	8
Negative	Positive	Sarcastic	5
Positive	Negative	Unclear sentiment indicator	19
Negative	Positive	Unclear sentiment indicator	2

Table 6: Results of labelling sarcasm, mixed emotions and unclear sentiment for misclassified instances.

Facing of emoticons: We therefore investigate another possible error source following (Mourad and Darwish, 2013), who claim that the right-to-left alphabetic writing of Arabic might result in emoticons being mistakenly interchanged while typing. On some Arabic keyboards, typing “)” will produce the opposite “ (” parentheses. The following example (2) illustrates a case of a misclassified instance, where we assume that the facing of emoticons might have been interchanged or mistyped.

(2) خلَاص مَافي امل :) *no hope anymore :)*

3.2 Lexicon-Based Distant Supervision

To avoid the issue of ambiguity in the direction of facing, we experiment with a lexicon-based approach to DS: instead of using emoticons, we now utilise the adjectives in our sentiment lexicon as noisy labels. We experiment with two different settings for the lexicon-based DS approach. First, we experiment with a lexicon-presence approach that automatically labels a tweet as a positive instance if it only includes positive lexicon(s) and the same for the negative class. Data instances having mixed positive and negative lexicons or no sentiment lexicons are excluded from the training set. The second approach is based on assigning a numerical value to sentiment words and aggregating the value into a single score, see Section 2. The results are summarised in Table 4.

Polar vs. neutral: We can observe that the models trained with the lexicon-presence approach significantly outperform the majority baseline, the fully supervised learning, as well as the lexicon-aggregation approach. The lexicon-presence and the emoticon-based DS approaches reach almost identical performance on our test set.

Positive vs. negative: Again, we observe that it is difficult to discriminate negative instances for both lexicon-based approaches. The lexicon-presence approach significantly outperforms the majority baseline, the fully supervised learning, and the lexicon-aggregation approach. But this time it also significantly outperforms the emoticon-based approach, which allows us to conclude that lexicon-based labelling introduces less noise for sentiment analysis. However, our results are significantly worse than the lexicon-based approach of Taboada et al. (2011), with up to 80% accuracy, and the learning-based approach of Zhanh et al. (2011), with up to 85% accuracy on English tweets. The lexicon-presence approach achieves the highest precision for negative tweets, see table 5, but still has a low recall. The lexicon-aggregation approach has the highest recall for negative tweets, but its precision is almost identical to the emoticon-based approach.

3.2.1 Error Analysis for Lexicon-Based DS

We conduct an error analysis in order to further investigate the difference in performance between the lexicon-presence and the lexicon-aggregation approach. We hypothesise that the lexicon-aggregation might perform better on instances with mixed emotions, i.e. tweets with positive and negative indicators, but a clear overall sentiment. We therefore manually add 36 instances to the test set which contain mixed emotions (but a unique sentiment label). However, the results on the new test set confirm the superiority of the lexicon-presence approach. In general, both lexicon-based approaches perform worse for sentiment classification. Taboada et al. (2011) highlight the issue of "positive bias" associated with lexicon-based approaches of sentiment analysis, as people tend to prefer using positive expressions and understate negative ones.

4 Conclusion and Future Work

We address the task of subjectivity and sentiment analysis (SSA) for Arabic Twitter feeds. We empirically investigate the performance of distant supervision (DS) approaches on a manually labelled independent test set, in comparison to a fully supervised baseline, trained on a manually labelled gold standard data set. Our experiments reveal:

(1) DS approaches to SSA for Arabic Twitter feeds show significantly higher performance in accuracy and F-score than a fully supervised approach. Despite providing noisy labels, they allow larger amounts of data to be rapidly annotated, and thus, can account for the topic shifts observed in social media.

(2) DS approaches which use a subjectivity lexicon for labelling outperform approaches using emoticon-based labels for sentiment analysis, but achieve a very similar performance for subjectivity detection. We hypothesise that this can be attributed to unclear facings of the emoticons.

(3) We also find that both our DS approaches achieve good results of up to 95% accuracy for subjectivity analysis, which is comparable to previous work on English tweets. However, we detect a decrease in performance for sentiment analysis, where negative instances repeatedly get misclassified as positive. We assume that this can be attributed to the more indirect ways adopted by people to express their emotions verbally via social media (Taboada et al., 2011). Other possible reasons for this, which we will explore in future work, include culturally specific differences (Hong et al., 2011), as well as pragmatic/ context-dependent aspects of opinion (Sayeed, 2013).

References

Muhammad Abdul-Mageed and Mona Diab. 2012. 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).

Muhammad Abdul-Mageed, Mona T. 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: short papers - Volume 2*, HLT '11, pages 587–591, Stroudsburg, PA, USA. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Sandra Kuebler, and Mona Diab. 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*, pages 19–28. Association for Computational Linguistics.

Rania Al-Sabbagh and Roxana Girju. 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).

Albert Bifet and Eibe Frank. 2010. Sentiment knowledge discovery in twitter streaming data. In *Discovery Science*, pages 1–15. Springer.

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

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

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12.

Lichan Hong, Gregorio Convertino, and Ed H Chi. 2011. Language matters in twitter: A large scale study. In *ICWSM*.

Ahmed Mourad and Kareem Darwish. 2013. Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. *WASSA 2013*, page 55.

Owen Rambow Nizar Habash and Ryan Roth. 2009. MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In Khalid Choukri and Bente Maegaard, editors, *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April. The MEDAR Consortium.

A. Pak and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREC*.

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 482–491, Avignon, France, April. Association for Computational Linguistics.

Jonathon Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, pages 43–48. Association for Computational Linguistics.

Eshrag Refaee and Verena Rieser. 2014. An Arabic twitter corpus for subjectivity and sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Asad Sayeed. 2013. An opinion about opinions about opinions: subjectivity and the aggregate reader. In *Proceedings of NAACL-HLT*, pages 691–696.

Jared Suttles and Nancy Ide. 2013. Distant supervision for emotion classification with discrete binary values. In *Computational Linguistics and Intelligent Text Processing*, pages 121–136. Springer.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 246–253, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Ian H Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Alecia Wolf. 2000. Emotional expression online: Gender differences in emoticon use. *CyberPsychology & Behavior*, 3(5):827–833.

Zheng Yuan and Matthew Purver. 2012. Predicting emotion labels for chinese microblog texts. In *Proceedings of the 1st International Workshop on Sentiment Discovery from Affective Data (SDAD)*, pages 40–47, Bristol, UK, September.

Omar F. Zaidan and Chris Callison-Burch. 2013. Arabic dialect identification. *Computational Linguistics*.

Ley Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. 2011. Combining lexiconbased and learning-based methods for twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011*, 89.