iLab-Edinburgh at SemEval-2016 Task 7: A Hybrid Approach for Determining Sentiment Intensity of Arabic Twitter Phrases

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

This paper describes the iLab-Edinburgh Sentiment Analysis system, winner of the Arabic Twitter Task 7 in SemEval-2016. The system employs a hybrid approach of supervised learning and rule-based methods to predict a sentiment intensity (SI) score for a given Arabic Twitter phrase. First, the supervised method uses an ensemble of trained linear regression models to produce an initial SI score for each given text instance. Second, the resulting SI score is adjusted using a set of rules that exploit a number of publicly available sentiment lexica. The system demonstrates strong results of 0.536 Kendall score, ranking top in this task.

1 Introduction

Sentiment Analysis (SA) concerns the automatic extraction and classification of sentiment-related information from a given text instance (Thelwall et al., 2012). This is the first time SA on Arabic text is considered in an international competition, like SemEval. Most of previous work on SA is in English, but there have been recent attempts to address SA for Arabic, e.g. (Abdul-Mageed et al., 2012; Mourad and Darwish, 2013; Refaee and Rieser, 2014c; Refaee and Rieser, 2014b). Previous work in this area has mainly focused on identifying the sentiment polarity in a given tweet/phrase, whereas within SemEval-2016 Task 7, the task is to predict the Sentiment Intensity (SI) in Arabic tweets. That is, in addition to their prior association to a sentiment class, i.e. positive or negative, each text instance has an SI score that indicates the strength of its assigned sentiment on a scale from 0 to 1.

In this work, we use a combination of supervised learning and rule-based approaches, exploiting a number of publicly available sentiment lexica. We find that the quality (rather than quantity) of these lexica influence system performance for the supervised part of the system. Our best performing system demonstrates strong results of 0.536 Kendall score, ranking top in SemEval-2016 Task 7. This type of hybrid approach between rule-based and statistical methods has been demonstrated to be successful in other shared tasks, such as dialogue state tracking (Wang and Lemon, 2013).

2 Related Work

Research on predicting Sentiment Intensity in Arabic is still limited. For example, El-Beltagy and Ali (2013) built a sentiment lexicon in which each entry is manually assigned an SI score. Using this lexicon, they calculated the overall Sentiment Orientation for a set of Egyptian tweets by adding up the score of extracted positive/negative words. The authors observed a significant improvement of up to 20.6% in accuracy when exploiting the SI scores, as compared to results using a uniform weighting scheme, i.e. positive word= +1 and negative word= -1.

A recent effort by Eskander and Rambow (2015) presents a large-scale sentiment lexicon for Arabic called SLSA wherein each entry is associated with an SI score. The scores are assigned using a linking algorithm that links the English gloss of each Arabic entry to a synset from SentiWordNet (Esuli and Sebastiani, 2006), which is a large-scale sentiment lexicon for English with SI scores. SLSA

Proceedings of SemEval-2016, pages 486–492,
San Diego, California, June 16-17, 2016.

is publicly available, and contains up-to-date coverage with nearly 35k lemma. However, SLSA covers only Modern Standard Arabic (MSA), which differs substantially from Dialectal Arabic (DA) typically used in social media platforms (Habash et al., 2013).

Other work that built sentiment lexica for Arabic either includes SA labels without SI scores, e.g. (Abdul-Mageed et al., 2011), or suffers from having duplicated and inflected (surface form) entries, e.g. (Abdul-Mageed and Diab, 2014). Others have not been made publicly available yet, e.g. (Mahyoub et al., 2014).

In this work, we make use of publicly available SI lexica and also contribute to ongoing efforts in automatically creating SI lexica for Arabic.

3 Approach

The proposed system uses a hybrid approach of supervised learning and rules for determining the sentiment orientation and assigning an SI score for a given Arabic phrase (see Figure 1). The assigned scores are real-valued ranging from 0 to 1, with the interval $[0, 0.5]$ associated with negative sentiment and the interval $[0.5, 1]$ associated with positive sentiment.

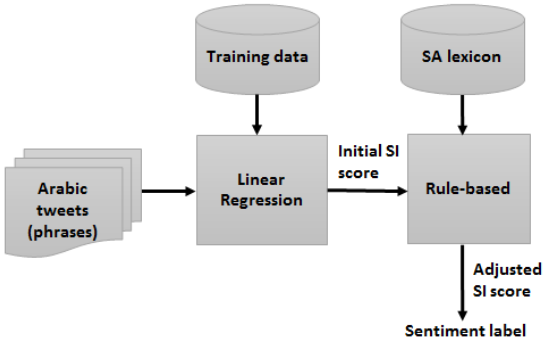

Figure 1: Hybrid system architecture.

3.1 Supervised Learning Component: Training LR models

The supervised part of the system uses Linear Regression (LR),[1] following Amir et al. (2015). To train the LR model, we use training data comprising publicly available sentiment lexica of positive/negative words along with their SI scores (training data-sets are described in section 4.1). We use word-lemma unigrams as features for training the LR models. The trained LR model is used to predict an initial SI score for each given text instance.

Training an LR model on thousands of data instances, such as those we used in our system to train the LR models (see section 4.1), can result in a significant increase in training time. For instance, we recorded a training time of more than 48 hours on a training-set of 10k instances (using a 64-bit operating system with 3.20 GHz, 48 core, 512GB RAM). Therefore, we experimented with several alternative settings. In our experiments, the best results (in terms of speed) are reached using a *bagging* method that generates multiple versions of a predictor, each of which is trained on a different sub-set of the learning data (Breiman, 1996). Each predictor produces a numerical value representing its prediction on a given test instance. The predictors are combined by averaging the output. In our experiments, we used an ensemble of 10 predictors, following (Banko and Brill, 2001), which produced a considerable reduction in training time (12.6 minutes to train an LR model using 10k instances).

3.2 Rule-based Component

In the second part of the system, the initial SI scores are passed to a rule-based component wherein a set of hand-crafted rules are applied to adjust the SI scores. The rules we use are inspired by those proposed by Taboada et al. (2011) and Thelwall et al. (2012) for lexicon-based SA. In particular, we use a combination of three publicly available sentiment lexica (see Section 4.2) [2] together with the following rules:

- Whenever a negative word from the combined lexicon (section 4.2) is detected, the SI score will be scaled towards negative $[0, 0.5]$.
- Same for positive words, except that the SI score will be scaled towards positive $[0.5, 1]$.
- If a negator is detected, the score will be shifted by a fixed amount, i.e. $+0.4/-0.4$. The

[1] We use WEKA's implementation of the LR scheme (Witten et al., 2013) with the default parameters configuration.

[2] Note that the sentiment lexica used in the 2nd phase are only associated with their sentiment labels, i.e. positive and negative, and are different from sentiment lexica used in the 1st phase of the system to train LR models.

only exception is when the SI score is between $[0.45, 0.55]$; then it will be considered neutral and will not be affected by negation.

- Finally, if no entries are found in the combined lexicon, then the final score will be the SI score initially assigned by the LR ensemble.

4 Data and Sentiment Lexica

4.1 Resources Used in the 1st Phase of the System: Training the LR Models

For the supervised LR models, we train with the following publicly available sentiment lexica that include SI scores. Note that the system that entered the competition only uses the labMT1.0 Sentiment Lexicon (see Section 6).

labMT1.0 Sentiment Lexicon: This is a compiled list of the most frequently-used 10k Arabic words from several resources including Twitter by Dodds et a. (2015). Each entry was manually annotated by 50 native speakers via Amazon MTurk using a nine-point-scale (1 very-negative/ 5 neutral/ 9 very-positive). We re-scale the manually assigned SI values to [0,1].

9k manually annotated Arabic Twitter data-set (Ar-tweet): This is a manually annotated and publicly available multi-dialect data-set of 9k Arabic tweets (Refaee and Rieser, 2014a). A feature vector representation of these tweets is created forming a list/lexicon of word-based unigrams. We add SI scores to this lexicon using an SVM classifier, following (Guyon et al., 2002). The classifier ranks the words according to how informative/useful they are for predicting the positive/negative label, see Table 1. Excluding words with a weight=0, the current list includes 9,785 words/features along with their weights/coefficients as assigned by the SVM classifier. Again, the SVM coefficients are re-scaled to [0,1].

SLSA v1.0 lexicon: This is a freely available sentiment lexicon for MSA (Eskander and Rambow, 2015). The lexicon is composed of nearly 35k entries annotated with their SI scores using a linking algorithm, as described in Section 2.

4.2 Resources Used in the 2nd Phase of the System: Rule-based Method

For the rule-based part of the system, the entries of the sentiment lexica do not need to be associated with SI scores. We therefore use a combination of the following resources:

ArabSenti sentiment lexicon: This is a freely available and manually annotated sentiment lexicon of 1,492 words that was created by Abdul-Mageed et al. (2011). Each entry is associated with a positive/negative sentiment label.

MPQA English sentiment lexicon: This is a manually annotated English lexicon that is created and made publicly available by Wilson et al. (2005). We automatically translate the lexicon (using Google Translate) and then manually filter it to remove irrelevant or no-sentiment-bearing words. The resultant lexicon includes 2,627 entries.

A manually annotated dialectal sentiment lexicon: This is a publicly available sentiment lexicon of 489 dialectal Arabic words. The lexicon is manually annotated by native speakers of Arabic (Refaee and Rieser, 2014a).[3]

4.3 Data Used for Developing and Evaluating the System

We use the data sets provided by SemEval Task 7.

SemEval'16 gold-standard development-set: This is a list of 200 instances (words/phrases taken from Arabic tweets) with their SI scores manually assigned. The entries can include negations. This set is used to evaluate different versions of the system.

SemEval'16 gold-standard test-set: This is a list of 1,166 instances (words/phrases taken from Arabic tweets) with their SI scores manually assigned. This data-set is used to evaluate the final system.

5 Data Pre-processing

We adopt a number of pre-processing techniques to tackle informality and alleviate the noise typically encountered in social media, following previous work, e.g. (Go et al., 2009; Bifet and Frank,

[3]The latter two lexica are available at: `http://goo.gl/qNLIZ2`

	Positive			Negative		
ID	**Arabic**	**English**	**SVM-weight**	**Arabic**	**English**	**SVM-weight**
1	مبروك	congratulations	0.7378	ابليس	devil	-0.0327
2	جميل	beautiful	0.6337	ارهَاب	terrorism	-0.5145
3	حلو	nice	0.5178	دمَار	destruction	-0.3653
4	ابدَاع	creative	0.4878	حقد	hatred	-0.3474
5	ابطَال	heros	0.0653	جحيم	hell	-0.3345

Table 1: Examples of the most predictive word uni-grams in the Ar-tweet data-set as evaluated by an SVM.

2010; Kouloumpis et al., 2011; Agarwal et al., 2011; Ahmed et al., 2013; Balahur et al., 2014; Rosenthal et al., 2014). The following procedures are applied to Ar-tweet (section 4.1) and SemEval's data-sets (section 4.3). Text lemmatisation is applied to all data described in section 4.

- **Normalising conventional symbols of Twitter:** this involves detecting entities like: #hashtags, @user-names, RT, and URLs; and replacing them by place-holders.
- **Normalising exchangeable Arabic letters**: mapping letters with various forms (i.e. *alef* and *yaa*) to their representative character.
- **Removing punctuations and normalising digits.**
- **Reducing emphasised words/expressive lengthening**: this involves normalising word-lengthening effects. In particular, a word that has a letter repeated subsequently more than 2 times will be reduced to 2 (e.g. *sadddd* is reduced to *sadd*).
- **Text lemmatisation:** we use lemmatised word-form to maximise coverage of the combined sentiment lexicon, following Taboada et al. (2011).[4]

6 Experiments and Results

We experimented using different combinations of lexical features (word-lemma unigrams) to train the LR models used in the 1st part of the system. The 2nd part is fixed throughout. Results are summarised in Table 2. The reported results are the final outcomes for the entire system, i.e. after adjusting IS scores in phase 2. Overall, we recorded an average improvement of 14% for applying the 2nd phase of the system. We report on Kendall's rank correlation coefficient (τ) and Spearman's coefficient (ρ) to account for SI ordering. Further details of the task, data and competing systems can be found in the task description paper (Kiritchenko et al., 2016).

System 1: (*official submission to SemEval-16)* This version of the system attained the best performance at a Kendall score of 0.5362 using lexical features based on the LabMT lexicon. This version has entered the competition and won Task 7.

System 2: This version uses features based on the Ar-tweet lexicon, which have resulted in a significantly ($p < 0.05$) lower performance as compared to LabMT at 0.0243 τ. A possible explanation for the performance variation is that the Ar-tweet lexicon is an auto-generated one. This auto-generation makes it prone to the inclusion of 'indirect' sentiment indicators (i.e. indicators which are merely inferred by the SVM model), because, for example, they are likely to appear in a negative political context. For instance, the SVM model assigned a strong negative weight of -0.78 for the feature *Bashar Al-Asad*, which is currently occurring in the context of the Syrian civil war. Thelwall et al. (2012) argue that such a feature can become outdated/irrelevant at a different point in time. Furthermore, human annotators, such as those recruited to annotate SemEval's test-set, are more likely to assign a neutral SI score to a feature like *Bashar Al-Asad*. In future, we will explore setting threshold values to filter features with lower SVM-weights as a mechanism to avoid the presence of indirect sentiment-bearing features.

System 3: Using SLSA, this version of the system is able to attain a comparable score to that recorded

[4]For lemmatisation, we use a stat-of-the-art Arabic morphological analyser, namely MADAMIRA v1.0 (Pasha et al., 2014).

System	Features	Kendall's τ coefficient	Spearman's ρ coefficient
1*	***labMT1.0***	**0.5362**	**0.67997**
2	Ar-tweet SVM-coeff.	0.0243	0.04329
3	SLSA	0.5244	0.65647
4	1 + 2	0.5261	0.66825
5	1 + 3	0.5256	0.67461
6	1 + 2 + 3	0.5141	0.66450

Table 2: Results on SemEval'16 gold-standard test-set using different lexical features for LR models. *System 1 entered the competition.

with LabMT. Although auto-generated (see section 4.1), SLSA is able to attain a Kendall score of up to 0.5244. SLSA has the advantage of being more than 3 times larger in size than LabMT (System 1) and Ar-tweet (System 2). In addition, the auto-generated SI scores in SLSA differs from the ones generated with SVM in Ar-tweet (System 2), as the former relies on linking Arabic entries to their corresponding English synset in SentiWordNet (Esuli and Sebastiani, 2006). Furthermore, being only based on MSA, SLSA can be assumed to be less noisy, especially when mapped to lemma-form, compared to slang and spelling variation in DA.

System 4: Combining LabMT and Ar-tweet has resulted in a comparable performance to System 3 at a Kendall score of 0.5261. However, this system is still not able to outperform that using LabMT on its own (see System 1). A possible explanation is that the presence of Ar-tweet results in introducing more noise than improving coverage of features (see System 2), resulting in slight degrading below the score attained by LabMT on its own (System 1).

System 5: Combining LabMT and SLSA has resulted in a slight improvement over using SLSA on its own, but still not competing with the performance of LabMT. However, when only considering phase 1 on its own, LabMT+SLSA performs best at a Kendall score of 0.377. In future work, we plan to investigate possible interactions between SLSA and the lexica used in phase 2.

System 6: Finally, combining all the training data still cannot reach the performance of LabMT on its own. It is also interesting to note that adding the auto-generated Ar-tweet caused a slight drop in Kendal score, compared to only using LabMT+SLSA (System 5).

7 Conclusion

This paper describes the iLab-Edinburgh Sentiment Analysis system, which is the top performing system of Arabic Twitter subtask for SemEval-2016 Task 7 (Kiritchenko et al., 2016). The aim of the task is to determine Sentiment Intensity for phrases taken from Arabic tweets. The proposed system consists of two phases: First, an ensemble of linear regression models are trained on lexicon-based word-lemma unigrams. Second, the SI scores are adjusted using a set of rules, leveraging pre-existing sentiment lexica. We experiment with different lexica for training the LR models. We find that the best results are attained using a manually annotated lexicon, labMT1.0 Sentiment Lexicon (Dodds et al., 2015). We also observe a drop in performance when adding features based on an auto-generated lexicon, which we attribute to noise. This highlights the need for high quality sentiment lexica for this task.

References

Muhammad Abdul-Mageed and Mona Diab. 2014. SANA: A large scale multi-genre, multi-dialect lexicon for Arabic subjectivity and sentiment analysis. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

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 Ap-*

proaches to Subjectivity and Sentiment Analysis, pages 19–28. Association for Computational Linguistics.

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38. Association for Computational Linguistics.

Shehab Ahmed, Michel Pasquier, and Ghassan Qadah. 2013. Key issues in conducting sentiment analysis on Arabic social media text. In *Innovations in Information Technology (IIT), 2013 9th International Conference on*, pages 72–77. IEEE.

Silvio Amir, Wang Ling, Ramón Astudillo, Bruno Martins, Mario J. Silva, and Isabel Trancoso. 2015. INESC-ID: A regression model for large scale twitter sentiment lexicon induction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 613–618, Denver, Colorado, June. Association for Computational Linguistics.

Alexandra Balahur, Rada Mihalcea, and Andrés Montoyo. 2014. Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications. *Computer Speech & Language*, 28(1):1–6.

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting on association for computational linguistics*, pages 26–33. Association for Computational Linguistics.

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

Leo Breiman. 1996. Bagging predictors. *Machine learning*, 24(2):123–140.

Peter Sheridan Dodds, Eric M. Clark, Suma Desu, Morgan R. Frank, Andrew J. Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M. Kloumann, James P. Bagrow, Karine Megerdoomian, Matthew T. McMahon, Brian F. Tivnan, and Christopher M. Danforth. 2015. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8):2389–2394.

Samhaa R El-Beltagy and Ahmad Ali. 2013. Open issues in the sentiment analysis of Arabic social media: A case study. In *Innovations in Information Technology (IIT), 2013 9th International Conference on*, pages 215–220. IEEE.

Ramy Eskander and Owen Rambow. 2015. SLSA: A sentiment lexicon for Standard Arabic. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2545–2550, Lisbon, Portugal, September. Association for Computational Linguistics.

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422.

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

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.

Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological analysis and disambiguation for dialectal Arabic. In *HLT-NAACL*, pages 426–432.

Svetlana Kiritchenko, Saif M. Mohammad, and Mohammad Salameh. 2016. Semeval-2016 task 7: Determining sentiment intensity of English and Arabic Phrases. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval'16, San Diego, California, June.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the OMG! *In Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 11:538–541.

Fawaz HH Mahyoub, Muazzam A Siddiqui, and Mohamed Y Dahab. 2014. Building an Arabic sentiment lexicon using semi-supervised learning. *Journal of King Saud University-Computer and Information Sciences*, 26(4):417–424.

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

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Eshrag Refaee and Verena Rieser. 2014a. An Arabic twitter corpus for subjectivity and sentiment analysis. In *9th International Conference on Language Resources and Evaluation (LREC'14)*.

Eshrag Refaee and Verena Rieser. 2014b. Evaluating distant supervision for subjectivity and sentiment analysis on Arabic twitter feeds. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*.

Eshrag Refaee and Verena Rieser. 2014c. Subjectivity and sentiment analysis of Arabic twitter feeds with

limited resources. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools (OSACT)*.

Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. *Proc. SemEval*.

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

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173.

Zhuoran Wang and Oliver Lemon. 2013. A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information. In *Proceedings of the SIGDIAL 2013 Conference*, page 423–432, Metz, France, August. ACL, Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

Ian H Witten, Eibe Frank, and Mark A Hall. 2013. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.