Benchmarking Machine Translated Sentiment Analysis for Arabic Tweets

Eshrag Refaee
Interaction Lab, Heriot-Watt University
EH144AS Edinburgh, UK
eaar1@hw.ac.uk

Verena Rieser
Interaction Lab, Heriot-Watt University
EH144AS Edinburgh, UK
v.t.rieser@hw.ac.uk

Abstract

Traditional approaches to Sentiment Analysis (SA) rely on large annotated data sets or wide-coverage sentiment lexica, and as such often perform poorly on under-resourced languages. This paper presents empirical evidence of an efficient SA approach using freely available machine translation (MT) systems to translate Arabic tweets to English, which we then label for sentiment using a state-of-the-art English SA system. We show that this approach significantly outperforms a number of standard approaches on a gold-standard held-out data set, and performs equally well compared to more cost-intense methods with 76% accuracy. This confirms MT-based SA as a cheap and effective alternative to building a fully fledged SA system when dealing with under-resourced languages.

Keywords: Sentiment Analysis, Arabic, Twitter, Machine Translation

1 Introduction

Over the past decade, there has been a growing interest in collecting, processing and analysing user-generated text from social media using Sentiment Analysis (SA). SA determines the *polarity* of a given text, i.e. whether its overall sentiment is *negative* or *positive*. While previous work on SA for English tweets reports an overall accuracy of 65-71% on average (Abbasi et al., 2014), recent studies investigating Arabic tweets only report accuracy scores ranging between 49-65% (Mourad and Darwish, 2013; Abdul-Mageed et al., 2012; Refaee and Rieser, 2014b). Arabic SA faces a number of challenges: first, Arabic used in social media is usually a mixture of Modern Standard Arabic (MSA) and one or more of its dialects (DAs). Standard toolkits for Natural Language Processing (NLP) mainly cover the former and perform poorly on the latter [1]. These tools are vital for the performance of machine learning (ML) approaches to Arabic SA: traditionally, ML approaches use a "bag of words" (BOW) model (e.g. Wilson et al. (2009)). However, for morphologically rich languages, such as Arabic, a mixture of stemmed tokens and morphological features have shown to outperform BOW approaches (Abdul-Mageed et al., 2011; Mourad and Darwish, 2013), accounting for the fact that Arabic contains a very large number of inflected words. In addition (or maybe as a result), there is much less interest from the research community in tackling the challenge of Arabic SA for social media. As such, there are much fewer open resources available, such as annotated data sets or sentiment lexica. We therefore explore an alternative approach to Arabic SA on social media, using off-the-shelf Machine Translation systems to translate Arabic tweets into English and then use a state-of-the-art sentiment classifier (Socher et al., 2013) to assign sentiment labels. To the best of our knowledge, this is the first study to measure the impact of automatically translated data on the accuracy of sentiment analysis of Arabic tweets. In particular, we address the following research questions:

1. How does off-the-shelf MT on Arabic social data influence SA performance?

[1] Please note the ongoing efforts on extending NLP tools to DAs (e.g. (Pasha et al., 2014; Salloum and Habash, 2012)).

Proceedings of NAACL-HLT 2015 Student Research Workshop (SRW), pages 71–78,
Denver, Colorado, June 1, 2015. ©2015 Association for Computational Linguistics

2. Can MT-based approaches be a viable alternative to improve sentiment classification performance on Arabic tweets?
3. Given the linguistic resources currently available for Arabic and its dialects, is it more effective to adapt an MT-based approach instead of building a new system from scratch?

2 Related Work

There are currently two main approaches to automatic sentiment analysis: using a sentiment lexicon or building a classifier using machine learning. Lexicon-based approaches, on the one hand, utilise sentiment lexica to retrieve and annotate sentiment bearing word tokens for their sentiment orientation and then utilise a set of rules to assign the overall sentiment label (Taboada et al., 2011). Machine Learning (ML) approaches, on the other hand, frequently make use of annotated data sets, to learn a statistical classifier (Mourad and Darwish, 2013; Abdul-Mageed et al., 2011; Wilson et al., 2009). These approaches gain high performance for English tweets: a benchmark test on commercial and freely-available SA tools report accuracy levels between 65% - 71% on English tweets (Abbasi et al., 2014).

For Arabic tweets, one of the best results for SA to date is reported in Mourad and Darwish (2013) with 72.5% accuracy using 10-fold-cross validation and SVM on a manually annotated data set (2300 tweets). However, this performance drops dramatically to 49.65% - 65.32% accuracy when testing an independent held-out set (Abdul-Mageed et al., 2012; Refaee and Rieser, 2014c). One possible explanation is the time-changing nature of twitter (Eisenstein, 2013): models trained on data collected at one point in time will not generalise to tweets collected at a later stage, due to changing topics and vocabulary. As such, current work investigates Distant Supervision (DS) to collect and annotate large data sets in order to train generalisable models (e.g. Go et al. (2009)). Recent work by Refaee and Rieser (2014b) has evaluated DS approaches on Arabic Tweets. They report accuracy scores of around 57% which significantly outperforms a majority baseline and a fully supervised ML approach, but it is still considerably lower than scores achieved on English tweets.

In the following, we compare these previous approaches to an approach using automatic Machine Translation (MT). So far, there is only limited evidence that this approach works for languages lack large SA training data-set, such as Arabic. Bautin et al. (2008) investigate MT to aggregate sentiment from multiple news documents written in a number of different languages. The authors argue that despite the difficulties associated with MT, e.g. information loss, the translated text still maintains a sufficient level of captured sentiments for their purposes. This work differs from our work in terms of domain and in measuring summary consistency rather than SA accuracy. Balahur and Turchi (2013) investigate the use of an MT system (Google) to translate an annotated corpus of English tweets into four European languages in order to obtain an annotated training set for learning a classifier. The authors report an accuracy score of 64.75% on the English held-out test set. For the other languages, reported accuracy scores ranged between 60 - 62%. Hence, they conclude that it is possible to obtain high quality training data using MT, which is an encouraging result to motivate our approach.

Wan (2009) proposes a co-training approach to tackle the lack of Chinese sentiment corpora by employing Google Translate as publicly available machine translation (MT) service to translate a set of annotated English reviews into Chinese. Using a held-out test set, the best reported accuracy score was at 81.3% with SVM on binary classification task: positive vs negative.

Our approach differs from the ones described, in that we use automatic MT to translate Arabic tweets into English and then perform SA using a state-of-the-art SA classifier for English (Socher et al., 2013). Most importantly, we empirically benchmark its performance towards previous SA approaches, including lexicon-based, fully supervised and distant supervision SA.

3 Experimental Setup

3.1 Data-set

We follow a similar approach to Refaee and Rieser (2014a) for collecting the held-out data set we use for benchmarking. First, we randomly retrieve

tweets from the Twitter public stream. We restrict the language of all retrieved tweets to Arabic by setting the language parameter to *ar*. The data-set was manually labeled with gold-standard sentiment orientation by two native speakers of Arabic, obtaining a Kappa score of 0.81, which indicates highly reliable annotations. Table 1 summarises the data set and its distribution of labels. For SA, we perform binary classification using *positive* and *negative* tweets. We apply a number of common preprocessing steps following Go et al. (2009) and Pak and Paroubek (2010) to account for noise introduced by Twitter. The data set will be released as part of this submission.

Sentiment	**Pos.**	**Neg.**	**Total**
no. of tweets	470	467	937
no. of tokens	4,516	5,794	10,310
no. of tok. types	2,664	3,200	5,864

Table 1: Evaluation data-set.

3.2 MT-based approach

In order to obtain the English translation of our Twitter data-set, we employ two common and freely-available MT systems: Google Translate and Microsoft Translator Service. We then use the Stanford Sentiment Classifier (SSC) developed by Socher et al. (2013) to automatically assign sentiment labels (positive, negative) to translated tweets. The classifier is based on a deep learning (DL) approach, using recursive neural models to capture syntactic dependencies and compositionality of sentiments. Socher et al. (2013) show that this model significantly outperforms previous standard models, such as Naïve Bayes (NB) and Support Vector Machines (SVM) with an accuracy score of 85.4% for binary classification (positive vs. negative) at sentence level [2]. The authors observe that the recursive models work well on shorter text while BOW features with NB and SVM perform well only on longer sentences. Using Socher et al. (2013)'s approach for directly training a sentiment classifier will require a larger training data-set, which is not available yet for Arabic [3].

3.3 Baseline Systems

We benchmark the MT-approach against three baseline systems representing current standard approaches to SA: a lexicon-based approach, a fully supervised machine learning approach and a distant supervision approach (also see Section 2). The **lexicon-based baseline** combines three sentiment lexica. We exploit two existing subjectivity lexica: a manually annotated Arabic subjectivity lexicon (Abdul-Mageed and Diab, 2012) and a publicly available English subjectivity lexicon, called MPQA (Wilson et al., 2009), which we automatically translate using Google Translate, following a similar technique to Mourad and Darwish (2013). The translated lexicon is manually corrected by removing translations with a no clear sentiment indicator [4]. This results in 2,627 translated instances after correction. We then construct a third dialectal lexicon of 484 words that we extract from an independent Twitter development set and manually annotate for sentiment. All lexica are 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 using MADAMIRA (Pasha et al., 2014). 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. 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

[2] SSC distinguishes between 5 sentiments, including very-positive, positive, neutral, negative, and very-negative. For our purposes, all very-positive and very-negative were mapped to the standard positive and negative classes.

[3] SSC was trained using a set of 215,154 unique and manually labeled phrases.

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

Metrics	Google-Trans.+DL		Microsoft-Trans.+DL		Lexicon-based		Distant Superv.		Fully-supervised	
	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.
precision	44.64	92.52	56.60	91.60	75.87	77.72	52.1	73.3	48.2	59.7
avg. precision	68.58		74.10		**76.79**		63.5		54.3	
recall	21.27	55.67	25.53	53.74	36.81	32.12	86.6	31.7	89.4	14.1
avg. recall	38.47		39.63		34.46		**57.1**		49.7	
F-score	28.81	69.52	35.19	67.74	49.57	45.45	65.1	44.2	0.627	22.8
avg. F-score	49.16		51.46		47.51		**53.9**		41.6	
accuracy	71.28		76.34		**76.72**		57.06		49.65	

Table 2: Benchmarking Arabic sentiment classification: results for positive vs. negative

represent mixed-sentiment instances with an even number of sentiment words.

The **fully-supervised ML baseline** uses a freely available corpus of gold-standard annotated Arabic tweets (Refaee and Rieser, 2014c) to train a classifier using word n-grams and SVMs (which we found to achieve the best performance amongst a number of other machine learning schemes we explored).

The **Distant Supervision (DS) baseline** uses lexicon-based annotation to create a training set of 134,069 automatically labeled tweets (using the approach we described for the lexicon-based baseline), where the identified sentiment-bearing words are replaced by place-holders to avoid bias. We then use these noisy sentiment labels to train a classifier using SVMs. Note that previous work has also experimented with emoticon-based DS, but has found that a lexicon-based DS approach leads to superior results (Refaee and Rieser, 2014b).

4 Experiment Results

Table 2 summarises the results for comparing the above baselines to our MT-based approaches (using Google and Microsoft MT), reporting on per-class and average recall, precision and F-measure. We also measure statistical significance by performing a planned comparison between the top-performing approaches (namely, the lexicon-based baseline and the two MT systems) using χ^2 with Bonferroni correction on binary accuracy values (see Table 3). We observe the following:

- In general, MT-based approaches reach a similar performance to the more resource-intense baseline systems. There is no significant distance in accuracy between the MT-based approaches and the overall best performing lexicon-based approach.
- Microsoft MT significantly outperforms Google MT for this task.
- Overall, the fully supervised baseline performs worst. A possible explanation for that is the time-changing nature of Twitter resulting in issues like topic-shift resulting in word token-based features being less effective in such a medium (Refaee and Rieser, 2014c).
- MT-based SA approaches in general have a problem of identifying positive tweets (low recall and precision), often misclassifying them as negative. The reverse it true for the DS and fully supervised baselines, which find it hard to identify negative tweets. This is in line with results reported by Refaee and Rieser (2014b) which evaluate DS approaches to Arabic SA. Only the lexicon-approach is balanced between the positive and negative class. Note that our ML baseline systems as well as the English SA classifier by Socher et al. (2013) are trained on balanced data sets, i.e. we can assume no prior bias towards one class.

Planned Contrasts	χ^2 (p)	Effect Size (p)
Google MT vs. Microsoft MT	273.67 (p=0.000)*	0.540 (p=0.000)*
Microsoft MT vs. lexicon-based	1.64 (p=0.206)	0.042 (p=0.200)
lexicon-based vs. Google MT	3.32 (p=0.077)	0.060 (p=0.068)

Table 3: Comparison between top approaches with respect to accuracy; * indicates a sig. difference at p<0.001

4.1 Error Analysis

The above results highlight the potential of an MT-based approach to SA for languages that lack a large

	Example Tweet	Human Translation	Auto Translation	Manual	Auto Label
1	ولي عهد بريطانيا طالع كشخه في الزي السعودي	Crown Prince of Britain looks very elegant in the Saudi attire	Crown Prince of Britain climber Kchkh in Saudi outfit	positive	negative
2	هذا الشبل من ذاك الاسد، الله يعافيك و يطول بعمرك	That cub is from that lion, God bless you with a healthy and long life	That drops of Assad God heal and go on your age	positive	negative
3	فرحّه محمد بالهدف	Muhammad's happiness with scoring a goal	Farahhh Muhammad goal	positive	negative
4	يا الله امطر اهل سوريا بالامن والرزق	Oh God, shower people of Syria with safety and livelihood	Oh God rained folks Syria security and livelihood	positive	negative
5	وعشان انكم معايا انا امتليت حياة، امتليت حب	Because you are with me, I'm full of life and love	And Ashan you having I Amtlat Amtlat love life	positive	negative
6	القمه الحكوميه في دبي بصراحه عمل يستحق التقدير، روعه	Frankly, the Government Summit in Dubai is a splended work that deserves recognition	Government summit in Dubai Frankly work deserves recognition, splendor	positive	negative

Table 4: Examples of misclassified tweets

training data-set annotated for sentiment analysis, such as Arabic. In the following, we conduct a detailed error analysis to fully understand the strength and weaknesses of this approach. First, we investigate the superior performance of Microsoft over Google MT by manually examining examples where Microsoft translated data is assigned the correct SA label, but the reverse is true for Google translated data, which is the case for 108 instances of our test set (11.5%). This analysis reveals that the main difference is the ability of Microsoft MT to maintain a better sentence structure (see Table 5).

For the following example-based error analysis of the MT approach, we therefore only consider examples where both MT systems lead to the same SA label, taking a random sample of 100 misclassified tweets. We observe the following cases of incorrectly classified tweets (see examples in Table 4):

1. Example 1 fails to translate the sentiment-bearing dialectical word, 'elegant', transcribing it as Kchkh but not translating it.
2. Incorrectly translated sentiment-bearing phrases/idioms, see e.g. *that cub is from that lion* in example 2.
3. Misspelled and hence incorrectly translated sentiment-bearing words in the original text, see example 3 'Farahhh' ('happpiness') with repeated letters. This problem is also highlighted by Abbasi et al. (2014) as one of challenges facing sentiment analysis for social networks.
4. Example 4 shows a correctly translated tweet, but with an incorrect sentiment label. We assume that this is a case of cultural differences: the phrase "oh God" can have a negative connotation in English (Strapparava et al., 2012). Note that the Stanford Sentiment classifier makes use of a manually labeled English sentiment phrase-based lexicon, which may introduce a cultural bias.
5. Example 5 represents a case of correctly translated sentiment-bearing words (love, life), but failed to translate surrounding text ('Ashan' and 'Amtlat'). Bautin et al. (2008) point out that this type of contextual information loss is one of the main challenges of MT-based SA.
6. Example 6 represents a case of a correctly translated tweet, but with an incorrectly assigned sentiment label. We assume that this is due to changes in sentence structure introduced by the MT system. Balahur and Turchi (2013) state that word ordering is one of the most prominent causes of SA misclassification. In order to confirm this hypothesis, we manually

corrected sentence structure before feeding it into the SA classifier. This approach led to the correct SA label, and thus, confirmed that the cause of the problem is word-ordering. Note that the Stanford SA system pays particular attention to sentence structure due to its "deep" architecture that adds to the model the feature of being sensitive to word ordering (Socher et al., 2013). In future work, we will verify this by comparing these results to other high performing English SA tools (see for example Abbasi et al. (2014)).

Example Tweet	تويتر مَااقدر اوصف شنَاعته
Google Trans.	I really appreciate what Twitter Describe the Hnaath
Microsoft Trans.	Twitter what I describe his ugliness
Human Trans.	I cannot describe how ugly is Twitter

Table 5: Example tweet along with its Google, Microsoft and human translations

In sum, one of the major challenges of this approach seems to be the use of Arabic dialects in social media, such as Twitter. In order to confirm this hypothesis, we automatically label Dialectal Arabic (DA) vs. Modern Standard Arabic (MSA) using AIDA (Elfardy et al., 2014) and analyse the performance of MT-based SA. The results in Fig. 1 show a significant correlation (Pearson, $p<0.05$) between language class and SA accuracy, with MSA outperforming DA. This confirms DA as a major source of error in the MT-based approach. Issues like dialectal variation and the vowel-free writing system still present a challenge to machine-translation (Zbib et al., 2012). This is especially true for tweets as they tend to be less formal resulting in issues like misspelling and individual spelling variations. However, with more resources being released for informal Arabic and Arabic dialects, e.g. (Cotterell and Callison-Burch, 2014; Refaee and Rieser, 2014a), we assume that off-the-shelf MT systems will improve their performance in the near future.

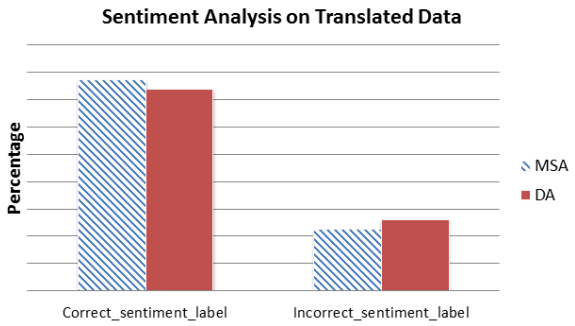

Figure 1: Performance of the sentiment classifier with respect to language class (MSA or DA)

5 Conclusion

This paper is the first to investigate and empirically evaluate the performance of Machine Translation (MT)-based Sentiment Analysis (SA) for Arabic Tweets. In particular, we make use of off-the-shelf MT tools, such as Google and Microsoft MT, to translate Arabic Tweets into English. We then use the Stanford Sentiment Classifier (Socher et al., 2013) to automatically assign sentiment labels (positive, negative) to translated tweets. In contrast to previous work, we benchmark this approach on a gold-standard test set of 937 manually annotated tweets and compare its performance to standard SA approaches, including lexicon-based, supervised and distant supervision approaches. We find that MT approaches reach a comparable performance or significantly outperform more resource-intense standard approaches. As such, we conclude that using off-the-shelf tools to perform SA for under-resourced languages, such as Arabic, is an effective and efficient alternative to building SA classifiers from scratch.

Future directions of this work include quantifying the impact of the used off-the-shelf tools, e.g. by using alternative high performing English SA tools. In addition, we plan to investigate multi-classifier systems, given the strength and weaknesses identified for each of the approaches.

References

Ahmed Abbasi, Ammar Hassan, and Milan Dhar. 2014. Benchmarking twitter sentiment analysis tools. In *Proceedings of the Ninth International Conference*

on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, may. European Language Resources Association (ELRA).

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.

Alexandra Balahur and Marco Turchi. 2013. Improving sentiment analysis in twitter using multilingual machine translated data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 49–55, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. 2008. International sentiment analysis for news and blogs. In *ICWSM*.

Ryan Cotterell and Chris Callison-Burch. 2014. A multi-dialect, multi-genre corpus of informal written Arabic. In *The 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland, May. European Language Resources Association.

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

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2014. AIDA: Identifying code switching in informal Arabic text. *EMNLP 2014*, page 94.

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

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

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

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)*.

Wael Salloum and Nizar Habash. 2012. Elissa: A dialectal to standard Arabic machine translation system. In *COLING (Demos)*, pages 385–392.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer.

Carlo Strapparava, Oliviero Stock, and Ilai Alon. 2012. Corpus-based explorations of affective load differences in arabic-hebrew-english. In *COLING (Posters)*, pages 1201–1208.

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

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 235–243. 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.

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

Makhoul, Omar F Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59. Association for Computational Linguistics.