OMAM at SemEval-2017 Task 4:
English Sentiment Analysis with Conditional Random Fields

Chukwuyem J. Onyibe and Nizar Habash
Computational Approaches to Modeling Language (CAMeL) Lab
Computer Science Program
New York University Abu Dhabi, Abu Dhabi, UAE
{chukwuyem.onyibe, nizar.habash}@nyu.edu

Abstract

We describe a supervised system that uses optimized Conditional Random Fields and lexical features to predict the sentiment of a tweet. The system was submitted to the English version of all subtasks in SemEval-2017 Task 4.

1 Introduction

Sentiment analysis, sometimes known as opinion mining, is the process of detecting the contextual polarity of text. That is, given a text (of any length), subjective information pertaining to the sentiment attached to the text is derived using natural language processing tools (Pang et al., 2008; Cambria et al., 2013). Sentiment analysis could be approached in two ways. General sentiment analysis, often termed sentence-level sentiment analysis, extracts the general sentiment of the text based solely on its contents. The sentiment is not related or based on any external entity. On the other hand, topic-level sentiment analysis infers the sentiment of the given text based on a specific topic. This branch of sentiment analysis has been further explored under the term Stance Detection (Faulkner, 2014; Anand et al., 2011). With the rapid increase in different forms of online expression like reviews, political criticism, ratings and punditry, social media has become an invaluable source of data for research in sentiment analysis. With data from social media, sentiment analysis can show the public sentiment towards current topics of public discourse. Twitter is one of the largest of such social media platforms and a prominent source of data for sentiment research (Pak and Paroubek, 2010; Wang et al., 2011; Rajadesingan and Liu, 2014). In this paper, we describe the components and results of a system for English sentiment analysis with which participated in an international shared task on sentiment analysis for Twitter data.

2 Shared Task Description

The SemEval-2017 Task 4 (Rosenthal et al., 2017) (henceforth SemEval) is aimed at categorizing tweets from Twitter. This task is composed of five subtasks. Subtask A is a message polarity classification task where tweets are classified on general sentiment (not directed at any topic) on a three-way scale: *Negative, Neutral and Positive* (henceforth, -1, 0, $+1$). Subtask B is a topic-based message polarity classification where tweets are classified on sentiment towards a given topic on a two-way scale: *Negative and Positive* (henceforth, -1, $+1$). Subtask C is the same topic-based task as B, except that it uses a five-point sentiment scale (-2, -1, 0, $+1$, $+2$), where -2 is *very negative* and $+2$ is *very positive*. Both subtasks D and E are tweet quantification tasks based on subtasks B and C, respectively. In both D and E, given the same datasets from B and C, the distribution of the tweets for each topic across each label of the given scales is estimated.

This task is a rerun of SemEval-2016 Task 4 (Nakov et al., 2016), with some changes. For this task, user profile information of the author of each tweet were made available. Also, this task included an Arabic language version. Our system works on English but is submitted as part of the OMAM (Opinion Mining for Arabic and More) team that also submitted a system that analyzes sentiment in Arabic (Baly et al., 2017).

3 Approach

For all subtasks, we used the same setup (process and system). We used CRF++ (Kudo, 2005), which is an implementation of Conditional Random Fields (CRF), as the underlying machine learning component. We were inspired by the work of Yang et al. (2007) who used CRFs to determine sentiment of web blogs, training at the sentence level and classifying at the document

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 670–674,
Vancouver, Canada, August 3 - 4, 2017. ©2017 Association for Computational Linguistics

level where the sentences sequence was taken in consideration. For this shared task, however, the tweets are not ordered, so there is no sequence information to be exploited. Nevertheless, we were interested in benchmarking how CRFs will fare in this scenario. We optimized the lexical features as well as the CRF++ parameters for each subtask independently against the specific subtask metrics. Although some subtasks involved topic-level sentiment analysis (i.e. sentiment towards a target), we ignored the topics for all subtasks. This idea is taken from the top scoring submission to SemEval-2016 Task 4C, TwiSE (Balikas and Amini, 2016), who used a single-label multi-class classifier and ignored topics altogether. For the tweet quantification tasks, we used a simple aggregation script (supplied by SemEval for a previous iteration of this task).

3.1 Data Preprocessing

We make use of all the data provided by SemEval for training and testing for all five subtasks. Additionally, we use a data set of 4,000 tweets available from SentiStrength.[1] In the SemEval data, each tweet is paired with a $TweetID$, $Topic$ and $Label$, except for subtask A data which has no $Topic$. For the SentiStrength data, each tweet is assigned two values representing positive and negative sentiment.

In order to use the training data from other subtasks and from SentiStrength in a subtask, we convert across the different label sets. For **subtask A** (three-point scale), we mapped subtask C's data's five-point scale labels -2 and $+2$ to -1 and $+1$, respectively; but used subtask B's data as is. We also added the SentiStrength data's two values and mapped them to $(-1, 0, +1)$. For **subtask B (and D)** (two-point scale), we folded the five-point labels as above and had two options regarding neutral values: remove the $neutral$ tweets or duplicate them and relabel them as $positive$ once and $negative$ once. We also explored classifying with higher point scales and mapping down. Details are discussed in Section 4.2. For **subtask C (and E)** (five-point scale), we converted data labels from other subtasks using duplication and relabeling: tweets with positive labels were duplicated and labeled with $+1$ and $+2$; tweets with negative labels were duplicated and labeled with -1 and -2; and the neutral labels (0) were simply duplicated to maintain balance. When converting subtask A data for use in other subtasks, a placeholder topic column was added. This did not influence the system as topics are not considered in any of the subtasks. The SentiStrength data was first mapped to subtask A format, then duplicated and relabeled.

3.2 Lexical Features

We considered the following lexical features with the CRF++ system. The unigram feature was always used, but feature combinations were explored for the other lexical features.

- **Unigrams** The unique words in each tweet consisting of alphanumeric characters and punctuation.
- **Tweet length (twtlen)** The number of words in the tweet.
- **Tweet length binned (twtlenbin)** The number of words in the tweet arbitrarily binned as LOW ($twtlen \leq 11$), MID ($12 \leq twtlen \leq 22$) and HIGH ($23 \leq twtlen$).
- **Bigrams** The unique bigrams in the tweet.
- **SentiStrength (senti)** The SentiStrength tool estimates the strength of positive and negative sentiment in short texts (Thelwall et al., 2010). The tool returns two values representing negative sentiment (range -1 to -5) and positive sentiment (range $+1$ to $+5$). Both values are used, as well as their sum, and a mapped value (onto the range of -2 to $+2$).
- **Removed URL (rurl)** All URLs are replaced with the string 'EXTERNALURL'. If the removed URL feature is $true$, that string is removed.
- **Stopwords (stpwrd)** This feature removes all stopwords in the tweet.

3.3 Model Optimization

We optimize the CRF++ model on the training data in two phases. First, the seven lexical features discussed above are exhaustively combined to identify the best feature combination for each subtask separately. Additionally, we explored combinations of different data sets, e.g., using SentiStrength data and/or subtask A data for subtask C. Using all of the available data for each subtask produced the best results. During this phase, the CRF++ is run with default parameter values. Next, the model is further optimized by tuning the CRF++ parameters c and f. The c value controls the hyper-parameter for the CRF to balance between overfitting and underfitting. The f parameter

[1] http://sentistrength.wlv.ac.uk/

Features	$AvgR$	$AvgF1$	Acc
unigram; senti	$\mathbf{0.623}$	$\mathbf{0.596}$	$\mathbf{0.686}$
unigram; twtlenbin; rurl; senti	0.621	0.593	0.667
unigram; stpwrd; twtlen; bigram; senti	0.613	0.584	0.667
unigram; twtlen; rurl; senti	0.613	0.583	0.686
unigram; twtlen; senti	0.613	0.580	0.688

Table 1: Scores of lexical feature combinations for subtask A

$-f$	$-c$	$AvgR$	$AvgF1$	Acc
$\mathbf{1}$	$\mathbf{8.5}$	$\mathbf{0.634}$	$\mathbf{0.612}$	$\mathbf{0.695}$
1	4.0	0.626	0.599	0.694
1	1.0	0.623	0.596	0.686
4	7.5	0.615	0.587	0.680
2	6.0	0.613	0.585	0.682

Table 2: Scores of CRF++ parameters combinations for subtask A

sets the cut-off threshold for the features. We explored all combinations of *c* and *f* ranging between 0.5 and 10.0 (in increments of 0.5) and 1 to 4 (in increments of 1), respectively.

3.4 Evaluation Metrics

Each subtask had its own target metric (Rosenthal et al., 2017). Subtasks A and B use $AvgR$, macro-averaged recall (recall averaged across the targeted labels). Subtask C uses MAE^M, macro-averaged mean absolute error. Subtask D uses KL, Kullback-Leibler Divergence. Subtask E uses EMD, Earth Mover's Distance.

4 Results

4.1 Subtask A

For subtask A, we used the following data sets for training: SemEval 2016 task 4A data (train, dev and devtest), SemEval 2016 task 4C data (train, test, dev and devtest) and SentiStrength twitter data. Table 1 shows the five top performing combinations from the lexical feature optimization phase. The *senti* feature with *unigrams* were the best features. Table 2 shows the five top performing combinations for the CRF parameter optimization phase. The best setup, with *c* value 8.5, *f* value 1 and features *unigram and senti*, was used to produce the predicted file submitted for subtask A for SemEval 2017. Our submission received the following scores and ranks (in subscript) out of 38 systems: average recall $AvgR$ = 0.590_{24}, $AvgF1$ = 0.542_{26}, *Acc* = 0.615_{19}.

4.2 Subtasks B and D

For subtasks B and D, we explored the possibility of training and predicting in five-point scale space and then mapping to two-point scale space. In the mapping, -2 and $+2$ map to -1 and $+1$, respectively. When the system predicts the $neutral$ label (0), we select the next most probable label determined using CRF++'s $verbose$ mode. For subtask B and D, we used the following training data: SemEval task 4B data, SemEval task 4A data, SemEval task 4C data and SentiStrength twitter data.

Table 3(a) shows the five top performing lexical feature combinations for subtask B. All features combinations in Table 3(a) were from the setup where classification is done in two-point format and the neutral tweets from subtask A and C data were removed.

Table 3(b) shows the five top performing c and f value combinations for subtask B. The highest performing setup, with c value 0.5 and f value 1 and features *twtlenbin, rurl, bigram and senti*, was used to produce the predicted file submitted for subtask B for SemEval 2017. Our submission received the following scores and ranks (in subscript) out of 23 systems: average recall $AvgR$ = 0.779_{15}, $AvgF1$ = 0.762_{17}, *Acc* = 0.764_{17}.

For subtask D, the five top performing feature combinations are shown in Table 5. All combinations in Table 5, except the second, are from the setup that predicts on five-point labeled data where the neutral tweets are preserved and duplicated. The second combination, *unigram; stpwrd; twtlen; senti*, is from the setup the predicts on five-point labeled data where the neutral tweets are removed.

The combination with the best score, used the following features: *unigram; twtlenbin, rurl, senti*. This was used for prediction for the test file which was later aggregated and submitted for subtask D. Our submission received the following scores and ranks (in subscript) out of 15 systems: KL = 0.164_{12}, *AE* = 0.204_{12}, RAE = 2.790_{12}.

Features	$AvgR$
unigram; twtlenbin; rurl; bigram; senti	**0.785**
unigram; stpwrd; twtlenbin; rurl; bigram; senti	0.783
unigram; twtlen; rurl; bigram; senti	0.783
unigram; stpwrd; rurl; bigram; senti	0.782
unigram; twtlenbin; bigram; senti	0.782

(a) lexical feature combinations

$-f$	$-c$	$AvgR$
1	**0.5**	**0.786**
2	0.5	0.786
1	2.0	0.786
1	2.0	0.785
3	1.0	0.782

(b) CRF++ parameters

Table 3: Optimization scores for subtask B

Features	MAE^M
unigram; twtlen; twtlenbin; rurl; senti	**0.74**
unigram; stpwrd; twtlen; rurl; senti	0.77
unigram; stpwrd; rurl; bigram; senti	0.77
unigram; stpwrd; twtlenbin; senti	0.77
unigram; senti	0.77

(a) lexical feature combinations

$-f$	$-c$	MAE^M
2	**10.0**	**0.71**
1	0.5	0.74
1	1.0	0.74
2	5.5	0.75
2	7.5	0.75

(b) CRF++ parameters

Table 4: Optimization scores for subtask C

Features	KL
unigram; twtlenbin; rurl; senti	**0.035**
unigram; stpwrd; twtlen; senti	0.036
unigram; rurl; bigram; senti	0.037
unigram; stpwrd; twtlenbin; rurl; senti	0.038
unigram; twtlenbin; rurl; senti	0.038

Table 5: Scores of lexical feature combinations for Subtask D

Features	EMD
unigram; twtlenbin; rurl; senti	**0.070**
unigram; twtlen; twtlenbin; senti	0.087
unigram; twtlenbin; rurl	0.087
unigram; senti	0.088
unigram; twtlen; twtlenbin; rurl; senti	0.088

Table 6: Scores of lexical feature combinations for Subtask E

4.3 Subtasks C and E

For subtask C and E, we used the following data for training: SemEval 2016 task4A data, SemEval 2016 task 4C data and SentiStrength twitter data. Table 4(a) shows the five top performing lexical feature combinations for subtask C. Table 4 shows the five top performing c and f value combinations for subtask C.

The highest performing setup, with c value 10.0 and f value 2 and features *unigram; twtlen, twtlenbin, rurl and senti*, was used to produce the prediction file submitted for subtask C for SemEval 2017. Our submission received the following scores and ranks (in subscript) out of 15 systems: $MAE^M = 0.895_{10}$, $MAE^{\mu} = 0.475_1$.

For subtask E, Table 6 shows the five top performing lexical combinations.

The combination with the best score, used the following features: *unigram twtlenbin, rurl, senti.* This was used for creating the prediction file which was later aggregated and submitted for subtask E. Our submission received the following scores and ranks (in subscript) out of 12 systems: $EMD = 0.350_7$.

5 Conclusion and Future Work

In this paper, we presented a system for English sentence-level sentiment analysis of twitter using CRF++ and optimized lexical features. We explore feature combinations and tune CRF++ parameters values to find the best setup for each subtask. Overall, the unigram and SentiStrength (*senti*) features were always present in the best performing setups for all subtasks. In all subtasks other than A, binned tweet length (*twtlenbin*) and removing URLs (*rurl*) consistently helped.

We used this system to participate in the SemEval-2017 Task 4. The system's performance was middle of the pack, which was accomplished while ignoring topics in the topic-level tasks.

In the future, we will explore more lexical features and other CRF and SVM implementations. We also look forward to applying the same setup to other languages.

References

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis*. Association for Computational Linguistics, pages 1–9.

Georgios Balikas and Massih-Reza Amini. 2016. Twise at semeval-2016 task 4: Twitter sentiment classification. *arXiv preprint arXiv:1606.04351* .

Ramy Baly, Gilbert Badaro, Ali Hamdi, Rawan Moukalled, Rita Aoun, Georges El-Khoury, Ahmad Al Sallab, Hazem Hajj, Nizar Habash, Khaled Shaban, and Wassim El-Hajj. 2017. Omam at semeval-2017 task 4: Evaluation of english state-of-the-art sentiment analysis models for arabic and a new topic-based model. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 601–608. http://www.aclweb.org/anthology/S17-2099.

Erik Cambria, Björn Schuller, Yunqing Xia, and Catherine Havasi. 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems* 28(2):15–21.

Adam Faulkner. 2014. Automated classification of stance in student essays: An approach using stance target information and the wikipedia link-based measure. *Science* 376(12):86.

Taku Kudo. 2005. Crf++: Yet another crf toolkit. *Software available at http://taku910.github.io/crfpp/* .

Preslav Nakov, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Fabrizio Sebastiani. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, San Diego, California, SemEval '16.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2(1–2):1–135.

Ashwin Rajadesingan and Huan Liu. 2014. Identifying users with opposing opinions in twitter debates. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer, pages 153–160.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 501–516. http://www.aclweb.org/anthology/S17-2088.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61(12):2544–2558.

Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. 2011. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, pages 1031–1040.

Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2007. Emotion classification using web blog corpora. In *Web Intelligence, IEEE/WIC/ACM International Conference on*. IEEE, pages 275–278.