Cross Corpus Emotion Classification Using Survey Data

Armin Seyeditabari[1], Sara Levens[3], Cherie Maestas[2], Samira Shaikh[1], James Igoe Walsh[2], Wlodek Zadrozny[1], Christine Danis[2] and Onah P. Thompson[2]

Abstract. Although semantic analysis and machine learning are becoming well established parts of Natural Language Processing (NLP), extraction of discrete emotions from text remains an under-developed area. Even less frequently do we see application of these technologies to open-ended survey questions in fields such as political science, psychology, public policy and sociology. In these domains, the need for more fined-grained emotion analysis of text responses has become apparent, particularly for assessing nuanced responses of the population to unexpected high impact events or incidents. Doing such assessments in real time is even more difficult. We report preliminary results on an ambitious attempt to perform a cross-corpus emotion classification that applies data gathered in one survey to text collected at a different time from different sources. This research is one step in a broader agenda to create new NLP methods to code large-scale text data from surveys and social media to improve studies of emotion contagion through social media networks. Our report is based on a medium-scale experiment from a survey conducted in the Fall of 2016 during a crisis event. Preliminary evidence suggests that with careful calibration of survey instruments, and proper understanding of natural language expressions (encoded as machine learning features), a transfer of classification code should be possible for some strongly expressed and potentially actionable emotions, like anger.

1 INTRODUCTION

Emotions are central to interpersonal communication, fostering rapid spread of ideas in social situations. As routes of interpersonal communication have rapidly expanded with increased social media options in the last few decades, it is critical to understand the role of emotions in communication and perhaps even more critically, how emotions can spread from one person to another to give rise to social action. Furthermore, social dynamics associated with communication and trust are likely to amplify emotion through the rapid spread of messages via social media channels (e.g. Facebook, Reddit, and newspaper comments). To examine these dynamics, we need to develop a method of analyzing event-specific social media text that can be deployed quickly during high impact crises. One key challenge in doing so, however, is to accurately identify discrete emotions in text, when the text uses context-specific language. Contextual expressions of emotions often contain metaphors, symbols, irony, or implicit language, making it difficult to classify with dictionaries. Further, using general sentiment dictionaries to identify text as positive or negative is insufficient for identifying the potential for social actions. Instead, it is essential to code for discrete emotion such as anger, an emotion linked to willingness to take risks and act to punish those deemed responsible (see Druckman and McDermott [1]). Anger is notably distinct from other negative emotions like fear which promote watchfulness and risk aversion (ibid.).

1.1 Goals of our research program

The goals of our research program are to create a methodology and a set of tools for real time or close to real time analysis of emotions expressed in social media in response to an event. We propose doing this based on the idea of cross-corpus classification of social media texts, that is, using tools developed in the context of prior events to analyze new events.

This requires an automated pipeline that is capable of recognizing specific emotions, and a method for fine tuning the parameters of such a pipeline in the context of a new event. In this scheme, we would begin with the text from a flash survey in response to the event, which would be used to fine tune the parameters of the previously built text analyzer, e.g. vocabulary and distribution of emotions in this new case. Since the flash survey contains only a few hundred to a couple of thousands short texts, it can be annotated by humans very fast (e.g. using Amazon Mechanical Turk).

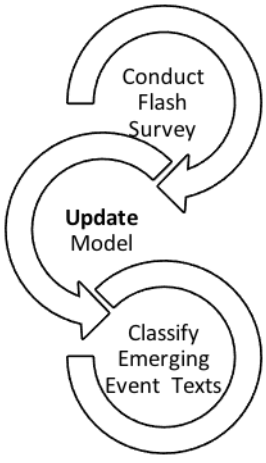

Figure 1. The pipeline for emotion classification using survey text: Flash surveys update parameters of the model, allowing real time understanding of emotions triggered by emerging events.

[1] Dept. of Computing Science, Univ. of North Carolina at Charlotte, USA. Email: {sseyedi1, wzadrozn, sshaikh2}@uncc.edu.

[2] Dept. of Political Science and public administration, Univ. of North Carolina at Charlotte, USA. Email: {cmaestas, jwalsh, cdanis1, pthomp34}@uncc.edu.

[3] Dept. of Psychological Science, Univ. of North Carolina at Charlotte, USA. Email: slevens@uncc.edu.

In this scheme, in the simplest case, we would begin with the text from a survey, and train a classifier using the preprocessed labeled text. This classifier would then use to identify emotional content in social media text (e.g. anger). The survey itself could then be kept in the field to collect small amounts of data as reactions to the incident changes pace or direction. Our classification model could then be updated reflecting these changes, and continuously used in analysis of the unfolding event.

1.2 Results in this paper

The goal of this paper is to describe our first experiments, and discuss the results from the perspective of cross-corpus classification, as presented above.

After a discussion of prior work on the use of natural language processing in analysis of emotions in text (Section 2) we describe in detail our approach.

The first step of it to conduct "flash" surveys dispersed via social media during and after a high-impact event; the surveys are asking respondents to describe their thoughts and feelings about the event in open-ended survey responses. The context-specific data is then extracted using human annotators, to be used in training of a machine learning classifier. (Section 3.1).

We pair the open-ended survey questions with closed-ended survey questions to explore whether using participants' self-reported emotions creates a more efficient means of training the model compared to human coding of the same text. (Section 3.2).

To achieve the objectives of our program, an important requirement is use a machine learning classifier. We analyze the performance of such classifiers based on the trained model generated from the human annotated flash survey text. This gives us the advantage of training our model with the language that is specific to the event. In this case, we analyzed the police shooting of Keith Lamont Scott in Charlotte, NC, which gained US-wide attention of all major media channels and social media.

We show that the classifier is capable of capturing the nuances in symbolic/metaphoric language that emerges as the meaning of the event is socially constructed (Section 3.3). We also describe multiple ways to build such models. Through this process, we demonstrate the advantages of this method, as well as some shortcomings and strategies to overcome them.

Finally, in Section 3.4 we describe our first experiments with cross-corpus classification, namely an application of the trained model to classify responses to the same event but in a different population and media outlets. These results are mixed; on the one hand we obtain high accuracy of 90.12% in finding expressions of anger; on the other hand, the classification of comments in news articles seems to be less reliable. We only have an anecdotal evidence for latter as in the over 450 comments we gathered, we have recognized automatically only a handful of 137 angry comments, probably due to the fact that most of them expressed anger only implicitly (Section 3.5).

2 OVERVIEW OF PRIOR WORK

The effect of emotion on human behavior has been the focus of many studies in a variety of social science and psychological disciplines. Recent research emphasizes the importance of studying the effects of distinct emotions rather than valence or sentiment (Lerner and Keltner [2]). Discrete emotions like anger and fear shape information search, cognition, motivation, and behavior in distinct ways (see [2,3,4]). The paper by Vasilopoulos [5] shows that "fear stemming from a terrorist attack will increase motivation to seek out political information, yet will have a negative effect on actual participation. In contrast, anger mobilizes participation in political action, even when such action entails an increased physical risk for the participant." Further, those who are angry tend to seek information related to blame and retribution, while those who are anxious seek information about protective measures [4]. Thus, emotion contagion in society is likely to produce different cascades of information depending upon whether the emotion that spreads is one of anxiety or anger. Studies such as these show the relevance of understanding emotions in political discourse and prediction of political events.

With the abundance of expressive text on the internet, especially in social media, natural language processing has been shown to be one of the most important tools in analyzing human behavior. In natural language processing, extraction of specific discrete emotions from text is still developing, and it benefits from prior work on sentiment analysis, (see [6] for discussion on how the two are related), sentiment and emotion extraction are distinct.

Dimensionality reduction methods to identify emotions are often used; e.g. Kim et al. [7] evaluate LSA, Probabilistic LSA and non-negative matrix factorization in identifying four emotions: Anger, Fear, Joy, and Sadness. Word embeddings are a newer promising methods of dimensionality reduction that build on distributional semantic models (see e.g. [8] for an overview of the distributional semantics promises and open problems). Although there has been some efforts in using word embeddings in semantic analysis (see [9,10,11]), Bellegarda's paper [8] is one of the few to consider word vector spaces for emotion analysis.

However, despite the growing presence of new methods of dimensionality reduction, so far most efforts in emotion extraction have been focused on classification using lexicon based techniques. In their paper, Staiano and Guerini [12] used crowd-sourcing to annotate around 37,000 terms with emotion scores to be used in emotion analysis tasks; Bobicev et al. [13] created a domain specific lexicon, HealthAffect, to analyze comments on an online health forum.

The use of statistical analysis methods also proved to be effective in many use cases. In [14], Vu et al. used a small sample of surveys about emotion provoking events to extract events with similar pattern from social media by seed expansion and clustering; and Kozareva et al. [15] used mutual information score to classify emotions in headlines. The most used method in literature for emotion extraction and analysis is the application of different classification techniques. For example, Yang et al. [16] showed that conditional random field classifiers outperform support vector machines in classifying emotions in web blog corpora; and Lin at al. [17] used classifiers with different feature settings to analyze emotions of readers of online news articles.

3 EXPERIMENT

Software As we described above, the broader goal of our research program is to create an automated pipeline, capable of

recognizing specific emotions in text and to apply it to coding social media text in near real time. In this section, we report preliminary experiments to test the efficacy of this approach. We focus on classifying one emotion in survey texts, namely anger. We focus anger because it serves as a catalyst for behavior (cf. [2] [3] [4]); thus, we wish to distinguish it from other emotions such as fear or frustration that are less likely to produce action.

3.1 Data and preparation

We compare two different ways in which event specific text could be manually coded to use in training a machine learning model for coding event-related social media text: (i) self-coding of emotion by survey respondents, and (ii) human coding by a researcher based on specific instructions.

We compare them in the context of a specific event, namely, the fatal police shooting of Keith Scott in Charlotte, North Carolina on 20 September 2016. This event occurred in the context of a national debate about the use of lethal force by police. The event attracted considerable attention from traditional and social media, because of Scott's race, ambiguity regarding important details, and the public release of police officer "body cam" video. For these reasons, the shooting generated a range of emotional responses.

Subjects in the Charlotte area were recruited to participate in the online survey via social media postings, postings of comments on local media websites, and from emails to the student population at a nearby public university. Participants were asked open-ended questions followed with close ended questions about the Keith Scott shooting. The main questions were focused on feelings and emotions about the incident and who or what were to blame.

After removing the empty answers from the survey, we had 1192 records, of which 839 were classified as angry. All texts were lowercased, and we used NLTK library in Python to remove punctuations and stop-words from the answers, and Porter stemmer in NLTK for stemming the words.

As we were primarily interested in the emotion classification, we mainly focused on the answers to the following open-ended, emotion question: "*Please tell us about the emotions you felt the most strongly in response to the news about the police shooting of Keith Scott and the protests that followed. Please tell us why you felt that way*." Next participants were asked to select all emotions they felt from the following options: angry, anxious, sad, disgusted, shocked, sympathetic, proud, frustrated, betrayed, and afraid and indicate to the strength or intensity of each selected emotion using a slider that ranged from 1 to 100. Anger was the focus of the present classification experiment, although the same process described here could be applied to any emotion-laden text.

3.2 Initial analysis

The results of using the close ended slider responses to label the open-ended emotion question, mentioned above, is presented below in Table 1. Two different classifiers, logistic regression and random forest, have been used with 10-fold cross validation. To assess the quality of the close-ended answers, we decided to use the slider responses (values between 1 to 100 selected by the participants) as the labels for our data (We chose 40% as the threshold as it showed to yield the best accuracy). The result of 10-fold cross validation on the balanced subset of our data did not show any improvement compared to the baseline ZeroR (majority class) classifier.

Baseline	Classifier	Accuracy	Recall
0.53	Logistic Regression	0.55	0.55
	Random Forest	0.61	0.61

Table 1. Using intensity slider as data label. Classifying answers to the emotion question using the emotion intensity slider (for anger) as label for the data. The low accuracy of the classifiers shows the inadequacy of using the emotion intensity sliders as label for the data.

This problem likely occurred because after answering the open-ended question in which the participants expressed emotion (e.g. disgust) they were presented with slider for a range of emotions and were asked to indicate how much they felt each emotion in response to the incident. Accordingly, even if they only wrote about one emotion in their written response, they could haven selected other emotions that they felt in response to the incident but did not write about. For example, one participant answered the emotion question like this:

"*I'm disgusted and saddened at the killing of Keith Scott and frustrated that there is nothing that I can do about it. The police were apparently serving a warrant, not intervening in the commission of a violent crime. I can't help thinking that they lost sight of what they were doing and severely discounted the value of Mr. Scott's life as they forgot their duty to protect. Nothing in any account I've seen or heard suggests that the police were in a situation where they had no choice. As simple and corny as it sounds, Andy Griffith would not have killed Keith Scott.*"

This participant's selected values for the emotion sliders was: 100 for angry, 40 for anxious, 81 for sad, 100 for disgusted, 50 for shocked, 49 for sympathetic, 69 for frustrated, and, 0 for betrayed, afraid and proud.

The failure of the direct attempt at using self-coded text for machine learning prompted us to try to use expert human labelling of the data, and to rethink the design of the survey for future research. In particular, future surveys deployed with this goal should better align close-ended responses and open-ended prompts so that the answers from close-ended questions could be used to label open ended responses, thereby creating an automated pipeline. To create the data to use in the human-coded experiments, we trained coders to recognize expressions of context-specific anger in survey text and hand code each survey response for the presence or absence of anger.

3.3 Building the model

After pre-processing the data as described in the previous section, we use Weka 3.8 [18] for our classifications. The StringtoWordVector filter in Weka was used to convert the document to feature vectors and all classifications was done with 10-fold cross validation. The baseline ZeroR classifier for this data was 70.38 percent accuracy.

For the first experiment, we ran the data through a logistic regression classifier, resulting in 83.7 percent accuracy in predictions with 0.72 recall. We also tried random forest with 83 percent accuracy and lower recall of 0.47. The resulting coefficients from logistic regression were used as measure for feature selection. After trying multiple values, we found that the margin of 2 for the coefficients results the best outcome for our classifier. We thus discarded the words with coefficients between -2 and 2, and with the resulting dataset we used a random forest classifier resulting in 91.23 percent accuracy with the recall of 0.79. The results are shown in Table 2.

Baseline	Classifier	Accuracy	Recall
0.70	Logistic regression (LR)	0.84	0.72
	Random forest (RF)	0.83	0.47
	LR (with feature selection)	0.81	0.70
	RF (with feature selection)	**0.91**	0.79

Table 2. Classifying answers to the emotion question using expert labels for anger. Feature selection has been done using coefficients from logistic regression. After feature selection using logistic regression, and using random forest as the classifier we reached a high accuracy of 91.23%.

We also tried using answers to two other survey questions as training data. The questions assessed feeling towards the protest:

"*We would like to know what you have felt or thought about the protests and riots that followed in the days after the shooting. Share the first things that come to mind*.", as well as who or what the participant blamed for the incident: "*Who do you think is most to blame for the protests and riots that occurred in Charlotte after the shooting*".

Of the survey's questions, we chose these two because the first one was one of the first open-ended questions on the survey which might have resulted in more expressive responses, and a natural association between anger and blame might support detection of anger text. Unfortunately, the outcomes of 10-fold cross validations (Table 3) revealed that the responses to these questions were poor training data. One potential reason for this is that due to the small size of the dataset is, more explicit and intense emotionally expressive language is needed to train the model.

For example, one participant whose answer to emotion question was labeled as angry:

"*Mad ! This makes Charlotte look like a circus. The protests SHOULD NOT be allowed . BLM should not be allowed to march the streets and preach hate! They are a awful group of angered individuals that act like the world is against them. They do not respect anyone but themselves. I'm just glad that I stayed away from it. I will NOT sit at a intersection and let someone disrespect me or tear up something I've worked hard for. They are heathens!*"

This person's answer to the feeling, and blame questions was respectively:

"*I think he was justified in his actions and should not be charged. They do not get paid enough for the job they are doing and I take it personally when they are disrespected."*

, and

"*Honestly the media. News channels and Facebook live. I watched protesters tell where they were and where they were going next. I watched thugs invite other thugs to join them!*"

One can see that although the anger is evident in the answer to the emotion question, no specific sign of anger is apparent in the language of the responses to the two other questions.

Baseline	Classifier	Accuracy	Recall
0.70	LR (question about blame)	0.56	0.37
	LR (question about feeling)	0.52	0.52
	RF (question about feeling)	0.68	0.68

Table 3. Classifying responses for anger based on questions about blame and feelings (using expert labelled data). The low accuracy of the classifiers shows that the answers to such questions are not a good source of training data.

3.4 Using new data

As our primary goal is to use the trained model from our survey to identify emotions from different sources of expressive text (e.g. social media, comments on news articles, etc.), we expanded the model to a naturalistic social media data set. This dataset was created by the same survey structure for a different population. This dataset contains 385 answers labeled by human experts. We used the previous dataset (the one with 1192 survey entries) as training data and used it to classify the new dataset.

The dataset was preprocessed in the same way as the first dataset, by lowercasing, removing punctuations and stop-words. Again, Weka was used to classify the data by using the first dataset as the training set, and the new dataset as the test set. The baseline accuracy for the test set was 71.73. After running the data through logistic regression for feature selection, and using random forest as our main classifier the resulting model had the accuracy of 90.12 percent with 0.90 recall. The summary of results can be seen in Table 4.

Our model was able to correctly classify the responses that was clearly angry, responses such as:

"*My strongest emotion was anger. Anger because the cops were being so poorly mistreated. I was also angry to see all the looting and assaulting of one another.*"

Also, it was able distinguish answers that clearly was showing other emotions like disappointment:

"*Disappointed that this is a problem. More than disappointed that some people fail to acknowledge that there is a disparity between how the black community is still treated and the white community*"

But in some cases, where the expressed emotions were a little ambiguous or misleading we could see misclassifications like the following response where many emotions were expressed:

"*First fear knowing people would go crazy. / Then a little hopeful that maybe since the officer was black people would wait until the facts were available. Of course then the family whips up emotions and withheld important facts. / Anger and fear as the events unfolded. / Disgust at people's behaviors towards the CMPD. / Great anger at random white people being targeted and attacked. Seems like no one wants to label that ad what it is: A Racist Hate Crime*."

Although anger was present in this answer, our model classified it as non-anger. Responses like the following which did not express anger, but include words that are mostly correlated with anger (e.g. fuels, anger, controversy) were misclassified as an angry comment:

"*The media fuels hate and controversy. They were showing all the anger the first 3 nights but not the majority that was peaceful. Everyone wants their 10minutes of fame and right now all you gotta do is be able to get ratings to have it.*"

Baseline	Classifier	Accuracy	Recall
0.72	LR	0.88	0.87
	RF	0.83	0.83
	LR (w/ feature selection)	0.77	0.77
	RF (w/ feature selection)	**0.901**	**0.90**

Table 4. Simplest case of cross-corpus classification. We are successfully classifying a new survey data using a machine learning model developed on another survey. High accuracy of 90.12% on test data shows the potential of this method for building a machine learning classifiers of expressed emotions, and using them to analyze new events.

3.5 Using online data

In order to see how this process would work on a totally new source of data, we attempted to classify news comments on an article about the same incident. We used 450 comments posted under the article "Amid Pressure, Charlotte Police Release Videos in Shooting of Keith Lamont Scott", taken from The Washington Post website. We used the same annotator who labeled 137 of those comments as angry. Using the same process of cleaning up the text, selecting features based on the results of logistic regression, and random forest as the classifier, the model classified most of the comments as non-anger. This showed that this model did not perform any better than a majority class classifier.

This experiment showed us the challenges in building the model based on the responses to the survey, and made us rethink the design of the survey so that the responses could better represent how people post comments on social media. To solve the issues that this experiment revealed we decided to make changes to future surveys and test the approach again. For example, in a recent survey we released, we tried to address these issues by changing the design of the survey to provide more focused questions asking about emotions expressed in the open-ended responses rather than asking a more general question about emotions felt in response to the event.

Solving this intermediate challenge is a necessary step towards empirically testing the implications of agent-based models of how emotional messages spread during crisis and spawn widespread social action.

4 CONCLUSION & FUTURE WORK

In this study, we showed the potentials and challenges in extracting emotions in social media posts about a specific incident. By conducting a flash survey about the event and collecting the emotional responses, we tried to classify emotions (e.g. anger) in online comments. Although we were able to create a model that could classify anger in a different survey population with high accuracy, the language difference in survey responses and social media posts, made it harder to use the model in real world situation. These results exposed the shortcomings of the current survey approach and showed us the ways we can change our design to overcome these shortcomings.

Creating a model that can classify emotions in social media posts based on a small sample of labelled text from a survey poses problems that need substantial effort to overcome. In this paper, our most successful classifiers were logistic regression and random forest; in another paper ([19]) we illustrate some of the challenges in using word embeddings to automatically reason about emotions in text. However, we haven't yet explored other successful classification techniques, e.g. ensemble methods ([20]).

Another problem we encountered was that the language and style between survey responses and social media posts differed considerably which may have hampered classification. The implicit nature of expressed emotion in an online comment versus explicit answers to a direct question about emotion in a survey was a significant obstacle and one of the main issues that caused this language difference, and made it hard for the classifier to distinguish the emotional text.

These experiments led us to change the design of our surveys in way that better aligns the test responses and the labelling of the data to be as close as possible to online posts. The many methods that we considered to achieve this goal are currently being implement in new surveys pertaining to high profile events such as the recent presidential inauguration and other emotionally charged political events.

We are now experimenting with new variants of survey questions that focus the self-coding on the emotions expressed in the text response and designing the open-ended survey questions to prompt for text they would post on a social media website.

Our overall findings are mixed. Trained expert coders can identify from survey responses the type of context specific anger expressions that can be used to classify text in a high profile event. However, the current attempts at cross-corpus classification using self-coded survey data was not successful. Trained human coding is effective, but does not meet our goal of developing a fully automated approach to machine learning for broader scale coding of survey and social media text during crisis events.

Although we see great potential to advance cross corpus emotion classification without the use of a human expert for labelling the survey answers, our initial tests direct attention to key considerations in the design of surveys to extract better quality text to use as training data. Extraction of discrete emotion from text is challenging because emotion expression is often context dependent, full of metaphors, symbols, and other implicit expressions. However, extant theory in social and psychological sciences point to the critical importance of identifying individual emotions because of their unique influences on information seeking, learning, attitudes, and behavior. Disentangling event specific emotion language in text is an essential step in understanding how social movements, unrest, protest, and crisis emerge and spread in response to critical events.

Our experiments showed us the challenging task of cross corpus classification, especially with a small training dataset. The small size of the dataset makes it hard to capture the intricacy of language and find features that can be used to classify new sources of text.

We have shown it is possible to successfully perform cross-corpus classification of free-text answers to a survey, and extract some (i.e. one, anger) of the emotions expressed in texts. We created a better method for constructing surveys that could be used in conjunction with automated classification tools. In our next steps, we will be simultaneously addressing the challenges of producing better automated analysis tools, extracting other emotions, and creating better survey designs.

REFERENCES

[1] J. N. Druckman and R. McDermott, "Emotion and the framing of risky choice," *Polit. Behav.*, vol. 30, no. 3, pp. 297–321, 2008.

[2] J. Lerner and D. Keltner, "Beyond valence: Toward a model of emotion-specific influences on judgement and choice" *Cognition and Emotion.* vol. 14, no. 4. pp. 473-493. 2000.

[3] B. Albertson, and S.K. Gadareian. *Anxious Politics: Democratic citizenship in a threatening world.* New York: Cambrdige University Press. 2015.

[4] T. J. Ryan. "What makes us click? Demonstrating incentives for angry discoures with digital-age field experiemtns" *The Journal of Politics.* vol. 74, no. 4, pp. 1138-1152. 2012.

[5] P. Vasilopoulos, "Terrorist Events, Emotional Reactions and Political Participation: Evidence from Two Studies Following the January and November 2015 Paris Attacks," *Emot. React. Polit. Particip. Evid. from Two Stud. Follow. January Novemb.*, 2015.

[6] A. Farzindar and D. Inkpen, "Natural language processing for social media," *Synth. Lect. Hum. Lang. Technol.*, vol. 8, no. 2, pp. 1–166, 2015.

[7] S. Mac Kim, A. Valitutti, and R. A. Calvo, "Evaluation of Unsupervised Emotion Models to Textual Affect Recognition," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 2010, pp. 62–70.

[8] J. R. Bellegarda, "Emotion analysis using latent affective folding and embedding," in *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, 2010, pp. 1–9.

[9] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 2011, pp. 142–150.

[10] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith, "Retrofitting word vectors to semantic lexicons," *arXiv Prepr. arXiv1411.4166*, 2014.

[11] D. Tang, F. Wei, B. Qin, T. Liu, and M. Zhou, "Coooolll: A deep learning system for twitter sentiment classification," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 2014, pp. 208–212.

[12] J. Staiano and M. Guerini, "DepecheMood: A lexicon for emotion analysis from crowd-annotated news," *arXiv Prepr. arXiv1405.1605*, 2014.

[13] V. Bobicev, M. Sokolova, and M. Oakes, "Recognition of sentiment sequences in online discussions," in *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, 2014, pp. 44–49.

[14] H. T. Vu, G. Neubig, S. Sakti, T. Toda, and S. Nakamura, "Acquiring a Dictionary of Emotion-Provoking Events.," in *EACL*, 2014, pp. 128–132.

[15] Z. Kozareva, B. Navarro, S. Vázquez, and A. Montoyo, "UA-ZBSA: a headline emotion classification through web information," in *Proceedings of the 4th international workshop on semantic evaluations*, 2007, pp. 334–337.

[16] C. Yang, K. H.-Y. Lin, and H.-H. Chen, "Emotion classification using web blog corpora," in *Web Intelligence, IEEE/WIC/ACM International Conference on*, 2007, pp. 275–278.

[17] K. H.-Y. Lin, C. Yang, and H.-H. Chen, "Emotion classification of online news articles from the reader's perspective," in *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*, 2008, vol. 1, pp. 220–226.

[18] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[19] A. Seyeditabari and W. Zadrozny. "Can Word Embeddings Help Find Latent Emotions in Text?". To appear in: *The 30th International FLAIRS Conference.* 2017.

[20] D. Opitz and R. Maclin. "Popular ensemble methods: An empirical study." *Journal of Artificial Intelligence Research* 11 (1999): 169-198