Are Emotions Detected Differently within First-Person Narratives?

Ryan Compton
University of California, Santa Cruz
Human Computer Interaction Lab
Santa Cruz, CA
rcompton@ucsc.edu

Abstract

Emotion classification within text has been attempted many times through lexical and implicit features. It remains to be seen if the extracted features from personal narratives can provide a base model for emotion classification. This paper explores the ability of traditional lexical and syntactic features used in previous emotion classification work toward personal narratives. The results are compared to a model developed from Non-Narratives and the results of Balahur et al. (2002), finding that Non-Narrative models produce the best results when tested within the Non-Narrative domain, but Narrative models produce the best generalized results.

1 Introduction

Text does not only communicate informative content, but also emotional states [1]. Emotion within language has been attempted to be extracted previously [1, 3, 6], however the use of emotion specifically within personal narratives, in its use and expression, is still being explored [5]. Within narrative, the agents tend to elicit emotions of a personal kind [7]. Narratives remain to be an untapped medium to supply base expressions of emotions. How emotion is conveyed through narrative could give a strong aspect of how people use empathy in order to bring across an idea. Examining expression of empathy has been generalized to more specific areas of interest, for example detecting emotions indicative of suicidal behavior [4] and implicit vs explicit emotional expression [3].

One of the goals of this paper is to examine how emotion can be detected from personal narratives. Personal narratives have been found within online blogs, where people are looking to find support during a given situation. Within these blogs are various topics of discussions of sexuality, family and friends, and fashion. Even though emotional understanding seems rather straightforward, the mechanism of how emotions are conveyed through text remains to be fully explored.

Furthermore, the second goal is to compare how emotions are detected within non-narratives. This provides a comparison for narrative emotion detection. Specifically this paper is attempting to answer these questions: Are people conveying emotions differently between personal narratives and non-narratives? Would emotional language use within narratives be a stronger indicator for emotional expression outside of narratives?

It is hypothesized that implicit emotion expression will be more prominent within narratives, coinciding with the results provided by Balahur et al. (2012). Emotions are also not expected to be expressed in similar forms as people use different forms of communication or representations in order to express themselves correctly [1,6]. Emotion words and phrases are expected to be used for creating empathy for the reader and if successful then it would bring the reader to a better implied understanding of the state of the narrator.

This paper will be comparing lexical indicators of emotion as well as parts-of-speech frequencies to the performance of Balahur et al.'s results examining implicit emotion expressions. It will also be comparing the differences in models trained on Narrative vs Non-Narrative data and evaluated on Narrative and Non-Narrative test sets. It is expected that models trained on

Narrative data will have better performance toward both Narrative and Non-Narrative test sets. Lastly, emotions are expected to be seen within all topics of discussion as variation across them does not seem to be enough to incite different emotions.

2 Related Work

Existing literature is reviewed about the main topics of this paper: (a) Emotional Language, (b) Models of Emotion Detection.

2.1 Emotional Language

Sentiment of text has been able to help determine the intent of writers as well as "their attitudes, evaluations, and inclinations with respect to various topics" [1]. Sentiment analysis typically focuses on classification of a binary distinction, negative or positive. These distinctions are associated synonymously with negative and positive emotions. This approach is rather lacking in detection of emotion type as well as intensity, due to the coarse binary distinction.

Aman and Szpakowicz address the task of classifying emotional expressions within text. They describe an emotion annotation task of identifying emotion category, emotion intensity and the words/phrases that indicate emotion in text. Using blogs, they were able to get an "emotional-rich" dataset that offered a variety of writing styles, choice, and arrangement of words and topics [1]. They began using seed words for six basic emotional categories (Happiness, Sadness, Anger, Disgust, Surprise, and Fear), and found blog posts containing one or more of these words. Next, using human annotators, each sentence was annotated with the appropriate emotion and its affective content. Their focus on sentence level instead of document level classification allows for a dynamic progression of emotions throughout text. An additional annotation made was the intensity of the emotional text (high, medium, low). Aman and Szpakoxicz found that annotator agreement varied depending on the emotional category. Happiness and Fear were the two with highest agreement (0.79,0.77 Cohen's kappa), while Surprise had the lowest (0.60 Cohen's kappa). For intensity, high levels of intensity had a 0.72 kappa level of agreement, while Medium and Low had much lower levels (Medium = 0.46, Low (0.37). Following Aman and Szpakoxicz, this paper will also examine seed words toward specific classes of emotions at the sentence level.

Mohammad and Turney also approach the problem of emotion detection, but instead intend to improve upon the issues of low-quality and low-size of emotion lexicons used. Using Amazon Mechanical Turk for annotation, they report high quality emotional word annotations by comparing the annotations received to existing gold standard data sets, WordNet Affect Lexicon and the General Inquirer.

Both of these works have used the WordNet Affect Lexicon as well as the General Inquirer for a gold standard of emotional text. Given that Mohammad and Turney's work to provide an enriched emotion lexicon, this paper will experiment with that lexicon on its ability to classify emotion within narratives.

2.2 Models of Emotion Detection

In order to classify whether a piece of text is conveying a specific emotion, statistical models are typically used in order to infer the probability that text is associated with an emotion or not. Here is where the work described in the previous section is put to use.

Aman and Szpakoxicz performed a classification experiment on their sentence level annotations. Using the annotations that had consensus between the annotators, they attempted to classify whether a sentence was emotional or not. The features they used were a bag-of-words providing by their annotators as well as the words from the General Inquirer and WordNet Affect Lexicon. They were able to achieve 73.89% accuracy of emotion vs non-emotion detection. This work does provide a fair base-line of comparison in which the paper will use for a general comparison of emotion detection.

Alm, Roth, and Sproat (2005) also explored the text-based emotion prediction problem by using supervised machine learning models. Their goal was similar to Aman and Szpakoxicz in that they were classifying the emotional affinity of sentences in the narrative domain of children's fairy tales. Their work presents a decent amount of feature exploration, including both semantic and lexical features, toward this classification problem. The authors were able to obtain an accuracy of 69% when using these features to classify neutral vs emotional sentences. While this work is preliminary, it provides a good feature base to start from for this project.

A comparative analysis by Balahur, Hermida, and Montoyo (2012) was motivated by the fact that most existing approaches to emotion

detection (Sentiment Analysis being specifically cited) are based on word-level analysis of text and mostly detect only explicit expressions of sentiment. They argue that in many cases, emotions are not expressed by using words with an affective meaning. Typically, writers are describing real-life situations where the reader is able to relate the situation to a specific emotion. The authors compare their purposed model based on the EmotiNet knowledge base, to other well-established methods for emotion detection and find that their approach has produced the best results. The authors produce stronger results in the field of emotion detection by experimenting with a higher number of classes within their model. They examine models ability to classify 7 different emotions (Anger, Fear, Disgust, Guilt, Joy, Sadness, and Shame). This paper will attempt to classify such a domain of emotions within narratives, as well as examine their model of implicit emotion detection.

Previous work using argument forums was conducted on distinguishing the differences between a factual based argument and a feeling based argument. This work was able to produce two types of models in order to predict the two types of arguments. Both classification and regression type models were created, with the regression model being more representative to the data. The work was rather low in scope as it only examined lexical features of the text, specifically lexical features provided by the Linguistical Inquiry and Word Count tool (LIWC). Furthermore, the representation of emotion under one class (Feeling) is not a fine grain measure of the varying types of emotion that exists. However, the results of factual vs. feeling based classification produced similar results to that of Aman and Szpakoxicz (73% accuracy). These preliminary results provide support that LIWC can be a significant indicator toward emotional presence. However it remains to be seen if LIWC is a good predictor of emotional type, specifically within narratives. This paper is attempting to push some of this work by introducing new feature sets and allow for the identification of emotional types and focus on the narrative aspect of writing.

3 Methods

Four experiments were run to explore the differences in data sets as well as the differences in feature set performance.

1. The first experiment is comparing all models' performance on the narrative and non-narratives data sets. Each model, after being trained on the narrative and non-narrative training sets, will test on both test sets separately.
2. The second experiment is examining the influence of up-sampling the training set to have an equal distribution of classes.
3. The third experiment is examining each feature set's ability toward the classification problem.
4. The fourth experiment trained a model on both the narrative and non-narrative training sets combined and testing on a combined data set as well.

3.1 Data

Data has been gathered from the online forum site teenhut.com. Posts on this site are meant for discussion and narratives and non-narratives have been found of people describing a past event. Within the data, emotional and non-emotional sentences will be annotated. 60 narratives and 60 non-narratives have been obtained, on average containing 13 sentences per post, totaling in 830 narrative and 861 non-narrative sentences, each acting as an instance for the models. An example of the data is:

*"I just had to **lie** to my best friend to stop her from **committing suicide**. She'd just been **dumped** because of her other ex-boyfriend, and now her 'friend' (who **hates** her just never admits it) tells her she likes said ex-boyfriend. Problem was she was the only one that didn't know. I was going to tell her but I didn't see her and I'd only known for a day. Then she said 'If you knew and didnt tell me I'd never talk to you-' and I'm the only one that she can talk to about the **depression** b/c I'd never judge her or spread it round. Now, I'm **scared**. If she finds out; she'd probably **kill** herself. I dont know what to do anymore :'(_____ The world can be a **Crazy** place but only if you're in it Alone"*

It is expected that words, such as the ones highlighted in bold above, will be strong indicators of emotion, thus motivating the use of lexical features. Each narrative was annotated based on the types of emotions being expressed. 6 categories of emotion were annotated along with a neutral class accounting for the possibility that no emotion is being conveyed. Sentences are then classified based on the emotional use and formatted in this manner:

("*I just had to lie to my best friend to stop her from committing suicide",* Sadness)
("*Now, I'm scared*", Fear)

Table 1 shows the class distributions for both data sets, where both have a high skew toward the neutral class. This large skew was the motivation for training set up sampling and experiment 2.

Neutral	587	Neutral	554
Anger	36	Anger	89
Disgust	4	Disgust	1
Fear	76	Fear	15
Happiness	49	Happiness	64
Sadness	107	Sadness	145
Surprise	3	Surprise	4

Table 1. Left: Class distribution for Narrative Data Set, Right: Class Distribution for Non-Narrative Data Set

Each data set was split into a 70-30 training-test set; where 6 fold cross validation is used on the training set to provide a development set.

3.2 Feature Sets

3.2.1 Emotion Lexicon

The emotion lexicon created by Mohammad and Turney will be used to provide the word frequencies for each emotion category within the lexicon. The emotion frequency for each sentence will be calculated through the following:

$$F_E = \sum_{w \in S} \begin{cases} w \in E & = 1 \\ & 0 \end{cases} \; / \; |S|$$

For each emotion E, the number of words within the sentence S that are associated with E within the emotion lexicon are counted, and then divide by the total number of words within the sentence. This provides a normalized frequency to the size of the sentence.

3.2.2 Linguistic Inquiry and Word Count (LIWC)

LIWC was developed by Pennebaker et al. (2001) to provide an efficient and effective method for studying the various emotional, cognitive, structural, and process components present in individuals' verbal and written speech samples. LIWC has been used in previous work [1], and is a good feature set to include in determining emotions as it is intended to study emotional components of writing.

3.2.3 Parts of Speech (POS)

Similar to the Emotion lexicon frequencies, the parts of speech frequencies will be provided for each sentence. The formula below describes how POS frequencies were found:

$$F_{POS} = \sum_{w \in S} \begin{cases} w \in POS & = 1 \\ & 0 \end{cases} \; / \; |POS|$$

The frequency is calculated by summing each type of POS within the sentence and then dividing that by the total number of POS within the sentence. POS is a common feature for NLP studies and has been used in various studies on emotion detection [1, 3, 6].

3.3 Evaluation

Each model and experiment will have the same evaluation metrics. For overall performance, Accuracy, ROC curve, F-score, Precision, Recall, True-Positives and False-Positives are used. To examine each model's ability to classify specific emotion classes, ROC curve, F-score, Precision, and Recall are used. One metric cannot serve the purpose of correctly evaluating a model's ability within a classification task, giving the motivation for such a wide array of evaluation metrics.

Narrative-Narrative	Accuracy	ROC	Precision	Recall	TP-Rate	FP-Rate	F-Measure
Naïve Bayes	53.25	0.748	0.617	0.533	0.533	0.217	0.564
RandomForest (1000 trees, 1 feature)	**63.22**	**0.775**	**0.693**	**0.632**	**0.632**	**0.577**	**0.509**
LogReg (B =1, C=30, E=0.02)	42.36	0.651	0.621	0.464	0.464	0.162	0.491
Ibk(K=5)	57.08	0.625	0.557	0.571	0.571	0.353	0.561
Perceptron	57.85	0.735	0.634	0.579	0.579	0.222	0.599
SVM	60.15	0.619	0.621	0.602	0.602	0.363	0.573
Narrative-NonNarrative							
Naïve Bayes	53.48	0.622	0.602	0.535	0.535	0.326	0.559
RandomForest (1000 trees, 1 feature)	**68.99**	**0.647**	**0.482**	**0.69**	**0.69**	**0.686**	**0.568**
LogReg (B =1, C=30, E=0.02)	55.4264	0.564	0.549	0.554	0.554	0.426	0.549
Ibk(K=5)	59.68	0.507	0.544	0.597	0.597	0.533	0.561
Perceptron	33.72	0.609	0.623	0.337	0.337	0.169	0.393
SVM	59.68	0.524	0.561	0.597	0.597	0.548	0.564
NonNarrative-NonNarrative							
Naïve Bayes	51.16	0.753	0.68	0.512	0.512	0.163	0.563
RandomForest (1000 trees, 1 feature)	**74.41**	**0.776**	**0.735**	**0.744**	**0.744**	**0.537**	**0.684**
LogReg (B =1, C=30, E=0.02)	48.44	0.641	0.651	0.484	0.484	0.202	0.529
Ibk(K=5)	58.14	0.637	0.641	0.581	0.581	0.291	0.603
Perceptron	56.2	0.698	0.675	0.562	0.562	0.205	0.595
SVM	42.25	0.625	0.645	0.422	0.422	0.172	0.47
NonNarrative-Narrative							
Naïve Bayes	39.84	0.611	0.541	0.398	0.398	0.233	0.452
RandomForest (1000 trees, 1 feature)	**59.77**	**0.699**	**0.374**	**0.598**	**0.598**	**0.61**	**0.46**
LogReg (B =1, C=30, E=0.02)	25.28	0.55	0.527	0.253	0.253	0.153	0.298
Ibk(K=5)	38.69	0.53	0.463	0.387	0.387	0.394	0.414
Perceptron	35.24	0.585	0.476	0.352	0.352	0.274	0.394
SVM	27.58	0.56	0.483	0.276	0.276	0.156	0.339

Table 2. Experiment 1-All models' overall results

Narrative-Narrative		Precision	Recall	F-Measure
RandomForest (1000 trees, 1 feature)	Anger	0.5	0.032	0.061
	Disgust	0	0	0
	Fear	0	0	0
	Happiness	1	0.059	0.111
	Sadness	1	0.063	0.118
	Surprise	0	0	0
NonNarrative-NonNarrative				
RandomForest (1000 trees, 1 feature)	Anger	0.5	0.091	0.154
	Disgust	0	0	0
	Fear	0.857	0.261	0.4
	Happiness	0.75	0.2	0.316
	Sadness	0.75	0.222	0.343
	Surprise	0	0	0

Table 3. Experiment 1-Random Forest performance on individual emotion categories

	Balahur Precision / Current Precision	Balahur Recall / Current Recall	Balahur F-Measure / Current F-Measure
Anger	0.610 / 0.500	0.284 / 0.091	0.154 / 0.388
Fear	0.712 / 0.857	0.33 / 0.261	0.451 / 0.400
Disgust	0.692 / 0.000	0.202 / 0.000	0.313 / 0.000
Happiness	0.895 / 0.750	0.218 / 0.200	0.351 / 0.316
Sadness	0.336 / 0.750	0.895 / 0.222	0.489 / 0.343

Table 4. Experiment 1-Comparison to Balahur et al.'s results

4 Results

4.1 Experiment 1

This experiment resulted in providing a single model that performed best for this task. Table 2 shows those results. The Random Forest algorithm performed best for the majority of evaluation metrics over all tasks. Examining its overall performance across the different train and test sets, Random Forest was performing best when trained on Non-Narrative data and then tested on Non-Narrative data as well. This was against the hypothesis that emotion classification would be best when examining Narratives. But the ability for the trained model to generalize to other test sets showed that better performance was found when trained on Narrative data and tested on Non-Narrative data, than the inverse.

To examine the model's ability to classify specific emotions, evaluation on each category is necessary and provided by Table 3. Due to this paper's focus on emotional classification, the neutral class is ignored within Table 3, however it was found to be the highest correctly classified category, which isn't surprising as the class distribution is heavily skewed toward the neutral class. Despite the strong skew, Random Forest is able to produce fair predictions for the Anger, Fear, and Happiness emotion categories, with all having at least 0.75 precision. This was only within the Non-Narrative trained and test sets. However, recall was rather low for all emotions indicating that the model had a high amount of false negatives.

Table 4 is a comparison of these results with the results reported from Balahur et al. These results perform roughly the same when compared to Balahur et al. especially within the Fear and Happiness category, where precision had the highest variance of about 0.14 between the two sets of results. For Sadness, this experiment was able to outperform Balahur in terms of precision, but in recall and F-measure, Balahur reports better performance.

4.2 Experiment 2

Experiment 2 shows the effects of up sampling the data within the training set to provide an equal class distribution. The best two algorithms from experiment 1 were used: Naïve Bayes and Random Forest. The results are shown in Table 5. Up sampling was found to actually hinder the performance of Naive Bayes by a significant amount, specifically within the Non-Narrative-Narrative train-test task by reducing accuracy and recall by about half.

As for Random Forest, up sampling provided mixed results. For the most part none of the changes were very large, looking at the Narrative training tasks and the Non-Narrative-Narrative train-test task, the performance for all evaluation metrics was at most a difference of 1.2%. While within the Non-Narrative-Narrative train-test task, there were varying changes to the evaluation metrics. Accuracy did improve by around 2% with up sampling, but ROC decreased by 0.03. This variation is also seen within the precision and recall metrics as precision decreases with up sampling but recall increases. Up sampling shows to produce no significant effects for the Random Forest model, but appears to hinder the Naïve Bayes model by providing strong biases to the training data.

4.3 Experiment 3

Random Forest models were created using each individual feature set to evaluate their performance. LIWC features were found to produce very similar results to the overall model in terms of accuracy and ROC, but actually had a higher precision score. This suggests LIWC to be the strongest feature set used for the overall model (Table 6). POS features also show good predictive ability for both tasks and actually outperform LIWC features within the Non-Narrative

Narrative-Narrative		Accuracy	ROC	Precision	Recall	TP-Rate	FP-Rate	F-Measure
No Up Sampling	Naïve Bayes	67.43	0.772	0.643	0.674	0.674	0.428	0.623
Up Sampling	Naïve Bayes	53.25	0.748	0.617	0.533	0.533	0.217	0.564
No Up Sampling	RandomForest	62.07	0.775	0.563	0.621	0.621	0.601	0.483
Up Sampling	RandomForest	63.22	0.775	0.693	0.632	0.632	0.577	0.509
Narrative-NonNarrative								
No Up Sampling	Naïve Bayes	65.5	0.559	0.547	0.655	0.655	0.605	0.586
Up Sampling	Naïve Bayes	53.48	0.622	0.602	0.535	0.535	0.326	0.559
No Up Sampling	RandomForest	69.37	0.647	0.481	0.694	0.694	0.694	0.568
Up Sampling	RandomForest	68.99	0.647	0.482	0.69	0.69	0.686	0.568
NonNarrative-NonNarrative								
No Up Sampling	Naïve Bayes	70.54	0.752	0.601	0.705	0.705	0.592	0.619
Up Sampling	Naïve Bayes	51.16	0.753	0.68	0.512	0.512	0.163	0.563
No Up Sampling	RandomForest	72.51	0.818	0.784	0.725	0.725	0.566	0.65
Up Sampling	RandomForest	74.41	0.776	0.735	0.744	0.744	0.537	0.684
NonNarrative-Narrative								
No Up Sampling	Naïve Bayes	61.3	0.697	0.535	0.613	0.613	0.504	0.523
Up Sampling	Naïve Bayes	39.84	0.611	0.541	0.398	0.398	0.233	0.452
No Up Sampling	RandomForest	60.53	0.682	0.374	0.605	0.605	0.614	0.462
Up Sampling	RandomForest	59.77	0.699	0.374	0.598	0.598	0.61	0.46

Table 5. Experiment 2-Effects of up sampling

Narrative-Narrative	**Accuracy**	**ROC**	**Precision**	**Recall**	**TP-Rate**	**FP-Rate**	**F-Measure**
All-Features	*63.22*	*0.775*	*0.693*	*0.632*	*0.632*	*0.577*	*0.509*
LIWC	**62.83**	**0.764**	**0.749**	**0.628**	**0.628**	**0.589**	**0.499**
POS	**60.53**	**0.596**	**0.507**	**0.605**	**0.605**	**0.581**	**0.48**
EmoLex	**35.24**	**0.61**	**0.563**	**0.352**	**0.352**	**0.233**	**0.433**
NonNarrative-NonNarrative	**Accuracy**	**ROC**	**Precision**	**Recall**	**TP-Rate**	**FP-Rate**	**F-Measure**
All-Features	*74.41*	*0.776*	*0.735*	*0.744*	*0.744*	*0.537*	*0.684*
LIWC	**68.6**	**0.662**	**0.48**	**0.686**	**0.686**	**0.694**	**0.565**
POS	**71.31**	**0.663**	**0.674**	**0.713**	**0.713**	**0.565**	**0.653**
EmoLex	**31**	**0.501**	**0.521**	**0.31**	**0.31**	**0.325**	**0.385**

Table 6. Experiment 3-Feature set performance

	Accuracy	ROC	Precision	Recall	TP-Rate	FP-Rate	F-Measure
All-Data	**67.24**	**0.762**	**0.66**	**0.672**	**0.672**	**0.582**	**0.581**
Narrative-Narrative	63.22	0.775	0.693	0.632	0.632	0.577	0.509
NonNarrative-NonNarrative	74.41	0.776	0.735	0.744	0.744	0.537	0.684

Table 7. Experiment 4-All data model

task, with a higher accuracy, ROC, precision, recall, and F-measure.

The biggest surprise was the emotion lexicon's inability to perform well on this task. Considering that it was designed to support the identification of lexical features toward these specific emotions, these results indicate that this lexicon's ability to generalize outside of its original context may not be strong.

When comparing LIWC's performance between Narrative and Non-Narrative, it appears that LIWC is obtaining better performance within the Narrative model as the ROC is roughly 0.1 higher than the Non-Narrative model.

4.4 Experiment 4

The last experiment was to examine the combination of both Narrative and Non-Narrative train and test sets combination performance. This experiment has a model that was trained on both the Narrative and Non-Narrative training sets and then evaluated on both test sets. Table 7 shows that this model was able to perform better in most evaluation metrics than that of the Narrative model, but is performing worse than the Non-Narrative model for all metrics.

The intention of this task was to create a more generalized model. This generalized model does perform better than the Non-Narrative model when tested on Narrative data and it performs as well as the Narrative model when tested on Non-Narrative data. This demonstrates the generative power of the Narrative model as it can perform as well as the generalized model, in a generalized test set.

5 Conclusion

The results shown are similar results found within Balahur et al. for certain emotional categories (Fear, Happiness, and Sadness). Using a Non-Narrative training set provided the best results toward a similar domain test set, however when testing on non-similar domain test sets, the Narrative trained model performed best. While the Narrative model isn't as powerful toward its own domain as the Non-Narrative model, the generative performance being similar to that of the overall trained model provides support for using Narrative features toward emotion classification.

As for the feature experiment, both LIWC and POS sets provided similar performance to that of the overall model. Seeing such simple frequency features provide a varying difference between the two datasets implies there exists a difference in the lexical and syntactic use between Narratives and Non-Narratives, especially shown from the LIWC feature set performing significantly better for Narratives. Lastly against expectations, the emotion lexicon provided by Mohammad & Turney did not have good performance for this task.

An interesting parallel to explore would be to compare this work to the annotation work by Aman and Szpakoxicz. They found that the inclusion of the intensity metric can also provide a richer dataset for modeling. Within Narratives this can be an important relation to the structure of the story. Is emotional intensity related to the temporal event structure within a Narrative?

These experiments were not without flaws as it is still strongly lexical based and as Balahur et al. argues, lexical based models do not capture the implicit expressions of emotions. These experiments were also lacking in strong datasets. Further work must expand on the number of instances for training and testing. This experiment could also improve upon annotations for emotions by using more annotators and stronger reliability checks.

6 References

1. Alm, C. O., Roth, D., & Sproat, R. (2005, October). Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 579-586). Association for Computational Linguistics.

2. Aman, S., & Szpakowicz, S. (2007, January). Identifying expressions of emotion in text. In *Text, Speech and Dialogue* (pp. 196-205). Springer Berlin Heidelberg.

3. Balahur, A., Hermida, J. M., & Montoyo, A. (2012). Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems*, *53*(4), 742-753.

4. Desmet, B., & Hoste, V. (2013). Emotion detection in suicide notes. *Expert Systems with Applications*, *40*(16), 6351-6358.

5. Hänninen, K. (2007). Perspectives on the narrative construction of emotions.*ELORE (ISSN 1456-3010)*, *14*, 1.

6. Mohammad, S. M., & Turney, P. D. (2010, June). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 26-34). Association for Computational Linguistics.

7. Oatley, Keith. Emotions and the Story Worlds of Fiction. Edited by M.C. Green, J.J. Strange, and T.C. Brock. Narrative Impact: Social and Cognitive Foundations, 2002, 39–69.

8. Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*, 2001.