QMUL at SemEval-2025 Task 11: Explicit Emotion Detection with EmoLex, Feature Engineering, and Threshold-Optimized Multi-Label Classification

Angeline Wang, Aditya Gupta, Iran R. Roman and Arkaitz Zubiaga
School of Electronic Engineering and Computer Science, Queen Mary University of London
{ec24817, ec24878, i.roman, a.zubiaga}@qmul.ac.uk

Abstract

SemEval 2025 Task 11 Track A explores the detection of multiple emotions in text samples. Our best model combined BERT (fine-tuned on an emotion dataset) predictions and engineered features with EmoLex words appended. Together, these were used as input to train a multi-layer perceptron. This achieved a final test set Macro F1 score of 0.56. Compared to only using BERT predictions, our system improves performance by 43.6%.

1 Introduction

SemEval 2025 Task 11 Track A is about determining what emotion most people will think is reflected in a short text snippet (Muhammad et al., 2025). This is about the perceived emotion by a reader, not about how someone is truly feeling. This is important because the individual's actual emotional state of being is difficult to define with absolute certainty (Van Woensel and Nevil, 2019; Wakefield, 2021).

The task consists in identifying the presence of five emotions, i.e. joy, sadness, fear, anger, and surprise. The main chellenges include varying lengths of texts, and imbalance of emotions. We approached this by stacking WordPiece tokenisation, preprocessing, EmoLex words and BERT predictions as features, which then we pass to MLPs. Upon quantitative evaluation on the development set, we found that using separate models for each emotion and dynamic thresholding based on each emotion was the most effective system. Our code is openly available[1].

2 Background and Related Work

Emotion analysis, a subfield of sentiment analysis, seeks to identify nuanced emotional states in text rather than broad polarity (Liu, 2012). Early work focused on lexicon-based methods, such as EmoLex (NRC Emotion Lexicon), which maps words to primary emotions and remains foundational for explicit emotion representation (Mohammad and Turney, 2013). While lexicons like EmoLex provide interpretability, their static nature struggles with contextual nuances. Mohammad and Kiritchenko (2018) showed emotion co-occurrence patterns (e.g., anger-disgust) to refine multi-label predictions, but their work relied on rigid lexicon counts rather than context-aware scoring.

The shift toward multi-label detection addresses the limitation of single-label classification, as text often expresses overlapping emotions (Wiebe et al., 2005). SemEval tasks have driven progress with top systems hybridizing lexicons and neural models. For example, Fersini et al. (2022) combined lexicon-derived features with BERT for multimodal classification, while Kumar et al. (2024) optimized thresholds for LLM-based emotion detection. Although weighted losses provide gains (Demszky et al., 2020), few studies address sensitivity—adjusting boundaries for imbalance.

Traditional approaches like CountVectorizer-based lexicon scoring (Mohammad et al., 2018) treat emotion-linked words equally, ignoring discriminative power. Recent advances in hybrid paradigms highlight the need for weighted lexical integration and threshold optimization. Our work bridges this by integrating EmoLex with BERT, using positive weight calculation to amplify discriminative terms (e.g., "devastated" for sadness) and emotion-specific threshold optimization to balance precision and recall—advancing methods from generic sentiment analysis (Liu, 2012) to nuanced multi-emotion detection.

3 Task and Dataset

3.1 SemEval 2025 Task 11 Track A

This year, the task involves the multi-class detection of five different perceived emotions, and we have chosen to explore English text only (Muham-

[1] http://github.com/angelinewang/semeval-task-11-track-a

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 959–964

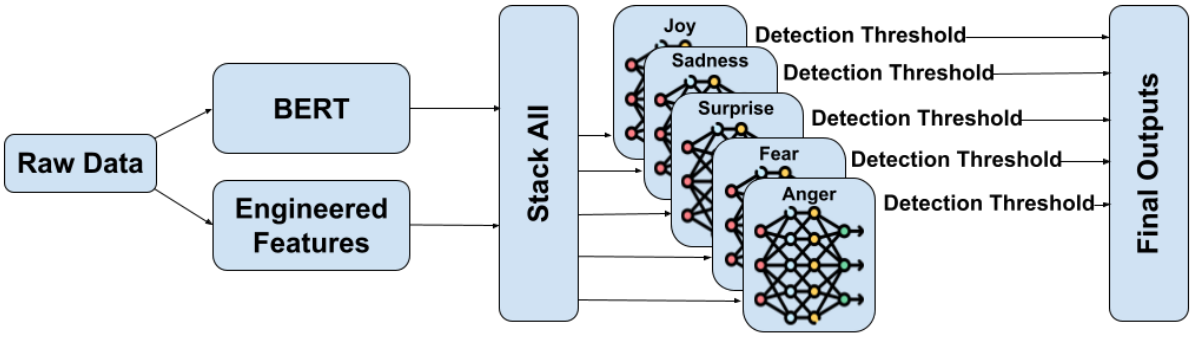


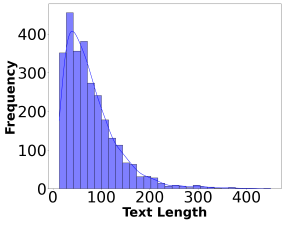
Figure 1: QMUL System Overview.

mad et al., 2025). This diverges from previous years' tasks, which focused on the speaker emotion (Kumar et al., 2024). This shift notably introduces cultural relativity due to annotators' diverse backgrounds, pragmatic ambiguity, and multi-perspective modeling (the need to predict majority perceptions rather than 'true' emotions).

3.2 Dataset

The dataset (Muhammad et al., 2025) includes 28 different languages, but we work only with English. Most of the data comes from social media posts (platforms such as Reddit, YouTube, Twitter, and Weibo). Some texts also include personal narratives, talks and speeches, which are anonymised. The data was human-annotated (through Amazon Mechanical Turk) by selection of all emotions applicable among five possible categories of perceived emotions: anger, sadness, fear, joy, and surprise. There was a total of 1222 annotators, and 5 to 30 annotators per sample. The training split has a size of 2,768, the development has 116, and the test set has 2,767.

The text length of the datasets follows a Zipfian Distribution as shown in the left panel of Figure 3. This is an important consideration due to the different context length constraints of different models.



Length	Train	Dev	Test
Avg	78	76	80
Min	13	9	5
Max	450	308	440
Range	437	299	435

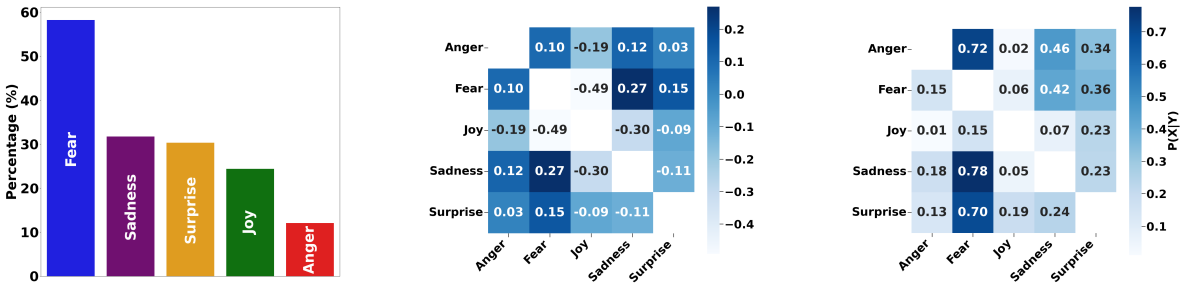
Figure 3: Left: Training set distribution of text lengths. Right: Statistics of text lengths per split.

3.3 Exploratory Data Analysis

Visualising the emotion class distribution that the dataset is imbalanced and that there is significantly more instances of 'Fear' and significantly fewer percentage of text had the presence of the emotion class 'Anger' in the dataset, as seen in Figure 2a.

To understand the relation between different emotion categories, we computed a correlation matrix (Figure 2b), which quantifies the co-occurrence tendencies of emotions across text samples. This showed that almost all correlations are statistically significant, with a p-value less than 0.05, excluding 'Anger' and 'Surprise'. Interestingly, 'Joy' is the only emotion that is anti-correlated with all other emotions. This means that the presence of 'Joy' is strongly indicative of the lack of the other emotion classes. Therefore, detecting different emotions with specific sensitivities (i.e. with Emotion-specific detection thresholds) is motivated by these distribution and correlation patterns.

Furthermore, Figure 2c shows the conditional probability matrix of the training dataset, with P(X|Y), where X and Y are emotion labels. This is important to see bidirectional relations between emotions. For example, fear has a large, often unidirectional association with many classes. The association is unidirectional, as it can be seen that given 'Fear', 'Anger' does not co-occur nearly as often. In contrast, the association between 'Fear' and 'Sadness' is somewhat more bidirectional, as given 'Fear', 'Sadness' occurs 42.3% of the time (Mohammad and Kiritchenko, 2018). 'Fear' is also the class with the most data samples in the training dataset; so this could just speak to the imbalance of data. One of this dataset's main purpose is to capture overlapping emotions, so it is natural that we see these co-occurrences, small and big, between emotion classes.



(a) Emotion Class Distribution (b) Class Correlation (c) Conditional Probability

Figure 2: Exploratory Data Analysis on Training Dataset.

4 System Overview

4.1 BERT Predictions as Features

As shown in Figure 1, we applied a pre-trained BERT (Savani, 2021; Devlin et al., 2018) model finetuned on an emotion dataset—dair-ai/emotion (Saravia et al., 2018) (this dataset included all five emotions under consideration, in addition to “love"). We chose to use the uncased version of BERT because it strips out accent markers and does not make distinctions between uppercase and lowercase letters. We used the model’s default tokenizer. And BERT was particularly useful for this task, as the texts are short in length (see Figure 3e). This means that BERT’s text length constraint does not interfere with its use for this case (Devlin et al., 2018).

We also chose to use the base variation of BERT because of our limited computational resources—Apple Mac laptops with M series chips—and the small size of the SemEval dataset (Muhammad et al., 2025).

We use BERT to obtain its class-wise emotion predictions for a given text d:

$$\mathbf{x}_{\text{bert}} = \text{BERT}(\text{Tokenize}(d)) \in \mathbb{R}^B, \qquad (1)$$

where $B = 5$ and $\mathbf{x}_{\text{bert}}$ contains the predicted probability distribution over 5 emotions.

4.2 Feature Engineering

In parallel to BERT predictions, the feature engineering pipeline includes: WordPiece tokeniser (with lowercasing, token for unknown words, separation token, padding token, classification token, mask token; tokens are further transformed into a numerical vector using CountVectorizer, which counts token frequencies), punctuation separation, stemming, lemmatization, bigram generation, and appending of matching EmoLex indicators.

After preprocessing, we check each token against the Emolex lexicon [2] to find ones that exists in EmoLex. This associates each token with one or more of the five target emotions. For each emotion class, a binary indicator is computed, and these five binary features are appended, yielding our "BoW representation". This step allows our model to learn from both contextual usage, through BERT, and explicit emotion associations, through EmoLex.

Therefore, document (i.e. a datapoint) d is represented as: $d \in \mathbb{R}^L$ where L = sequence length. This can be tokenized using the continuous bag of words CountVectorizer, thus yielding $\mathbf{x}_{\text{bow}} \in \mathcal{R}^V$, where V = vocabulary size (4,340 unique tokens).

4.3 Stacking All Features

Features from BERT predictions and feature engineering—capturing the full text contextual information—were stacked.

$$\mathbf{h}_0 = [\mathbf{x}_{\text{bow}} \| \mathbf{x}_{\text{bert}}] \in \mathbb{R}^D, \qquad (2)$$

where $x_{\text{bow}} \in \mathbb{R}^V$ is the BoW feature vector ($V = 4,340$) and $x_{\text{bert}} \in \mathbb{R}^B$ is the BERT output ($B = 5$, one per emotion class), yielding the final feature dimensionality of $D = V + B = 4,345$.

4.4 Emotion Detection MLPs

We used five multi-layer perceptron (MLP) classifiers, one for each emotion. Each MLP consists of a sequential arrangement of layers that process the stacked input of size D features in parallel.

The input layer reduces the dimensionality of the input to a lower dimensional space of 256, followed by batch normalisation and a ReLU activation. There are two subsequent projections. One

[2] EmoLex (Mohammad and Turney, 2013) maps frequent English words to explicitly emotion associations, providing interpretable signals.

that projects to a dimension of 128, and another one to a space of 64, hence further reducing dimensionality. Both of these projections also use batch normalisation and ReLU. The output layer condenses the features to a single output neuron that represents the probability of an emotion's presence, thus allowing for threshold-based detection. MLP learning is supervised using Binary Cross-Entropy Loss (Equations 3, 4).

$$L = \begin{cases} -w_p \text{BCE}(\hat{y}, y), & \text{if } y_p = 1 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

We add a *positive weight calculation* to handle class imbalance in the binary classification task. This modifies the loss function to penalise mistakes on the minority class more heavily:

$$w_p = \frac{\text{number of negative samples}}{\text{number of positive samples}} = \frac{N - \sum_i y_i}{\sum_i y_i} \quad (4)$$

where N is the total number of samples, $\sum_i y_i$ is the total number of positive samples (i.e., samples where $y_i = 1$). and $N - \sum_i y_i$ is the total number of negative samples (i.e., samples where $y_i = 0$).

4.5 Detection Threshold Selection

Emotion detection thresholds were selected based on the development set by optimising the F1 scores for each individual emotion. The thresholds found were used for the final predictions on the test set (Joy: 0.45, Sadness: 0.55, Surprise: 0.20, Fear: 0.50, Anger: 0.60). Our original motivation for looking for thresholds was due to the high imbalance in the proportion of data for each emotion, this can be seen in Figure 4a, which shows the final thresholds chosen with the proportion of the emotion data in the train set. The right panel of Figure 4b shows how each threshold impacts the Macro F1 score in the development set.

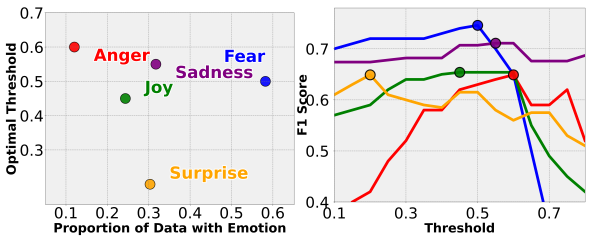


Figure 4: Left: Optimal thresholds for emotion detection aligned with the proportion of data for each emotion in the training set. Right: Variations in F1 score by detection threshold for each emotion. Colours in the left graph correspond to the right graph.

5 Experimental Setup

Input features. We assessed performance with three different variations of input features: (i) with inputs that consisted of BERT prediction logits only, (ii) with BERT predictions stacked on top of the engineered features, and (iii) with the addition of EmoLex words appended to the text that went into engineered features.

Classifier configurations. For each of the three variations above, we evaluate three dfiferent configurations of MLP classifiers: (i) a single MLP with a threshold of 0.5, (ii) a different MLP for the detection of each emotion, with a threshold of 0.5, and (iii) due to the imbalance of emotion data (Figure 2a), we assessed whether it might be useful to set differing detection thresholds for each emotion on top of just emotion-specific MLPs. These configurations help us evaluate the extent to which having different classifiers for each emotion helps.

Hyperparameters. All experiments used Adam (lr=$1e^{-3}$, weight decay=$1e^{-4}$), batch size 32, dropout (0.3/0.2), and a learning rate scheduler (patience=5, factor=0.5). This means that the learning rate is cut in half if there is no improvement of Macro F1 on the validation set for 5 epochs. Training was done for 400 epochs, with an early stopping mechanism with a patience of 10 epochs.

Evaluation. Performance was measured using the Macro F1 score.

	Dev	Test
Single MLP (0.5 Thrs)		
BERT	0.6080	0.5418
BERT+Feat ENG	0.6114	0.5624
BERT+Feat ENG+EmoLex	0.6075	**0.5638**
Emotion-Specific MLP (0.5 Thrs)		
BERT	0.6162	0.5434
BERT+Feat ENG	0.6476	0.5457
BERT+Feat ENG+EmoLex	0.6491	0.5538
Emotion-Specific MLP (Emotion-Spec Thrs)		
BERT	0.6568	0.5364
BERT+Feat ENG	0.6816	0.5542
BERT+Feat ENG+EmoLex	0.6433	*0.5558*

Table 1: Comparison of Macro F1 performance scores.

6 Results

For experiments with a single MLP and a 0.5 emotion detection threshold across all emotions, using

only BERT predictions as input features achieved a Macro F1 score of 0.6080 on the development set and 0.5418 on the test set. Using stacked engineered features on top of BERT predictions achieved an improvement leading to a Macro F1 of 0.6114 on the development set and 0.5624 on the test set. The addition was EmoLex words to the engineered features led to a drop in Macro F1 on the development set with a score of 0.6075, but an increase in test set performance to 0.5638—this ended up being our best model over all others on test set.

The variant with emotion-specific MLPs (but still with 0.5 detection thresholds), using only BERT predictions as input led to a model with a Macro F1 of 0.6162 on the development set (beating the performance of all preceding models) and 0.5434 on the test set. Stacking engineered features further increased the Macro F1 on the development set to 0.6476, and led to the a test set performance of 0.5457. Appending EmoLex words to the engineered features further increased the Macro F1 on the development set to 0.6491, and led to a test set performance of 0.5538.

Finally, using emotion-specific detection thresholds for each emotion-specific MLPs, when using BERT predictions as input, the development set performance continued to increase to 0.6568, but test set performance went down to 0.5364. Using engineered features in addition to BERT predictions led to an increase in development set performance with Macro F1 of 0.6816, with test set performance hovering at 0.5542. Finally, appending the EmoLex words to the engineered features led to a dip in development set Macro F1 performance to 0.6433, with a corresponding test set performance of 0.5558.

In general, we see that engineered features (without EmoLex words) always improved the performance on both the development and test set for all variations. On the test set, adding EmoLex words also consistently improved the performance of our models across all variations. On the other hand, EmoLex words resulted in a decrease in development set Macro F1 performance with emotion-specific MLPs and emotion-specific detection thresholds. It seems that EmoLex either does not create an impact or creates a slight negative impact when looking at development set performance. This shows that EmoLex words were non-specific to the validation set but allowed for better generalisability to unseen test data, which presumably included more of these words, and were perhaps indicative of classifications, which was maybe not the case for the smaller validation set. It is worth remembering that the test set is more than 20 times bigger than the development set. Thus, the size of the test set makes it more complex, challenging and thorough than the validation set, hence the generalisation of best models based on development set performance is not good.

Class-wise performance correlated with label frequency: the majority class (Fear) had the highest F1, while the rarest (Anger) showed lower recall (sparse training data). Joy performed well (moderate frequency), likely aided by anti-correlation with other emotions. Threshold optimisation partially addressed imbalance, but minority classes still lagged due to limited training data.

In summary, the results (Table 1) show that, in the development set, the best model included emotion-specific MLPs, emotion-specific thresholds, BERT, engineered features and EmoLex words. However, the best-performing model on the test set was BERT with the engineered features, EmoLex words, a single MLP and a 0.5 detection threshold for all emotions. This model ended up generalising the best, and other models that performed well on the development set tended to suffer heavily from domain shift when evaluated on the test set.

7 Conclusion

Our best model combined BERT predictions and engineered features (including EmoLex), which were used as input to an MLP. The detection was optimal using a 0.5 threshold, achieving a test Macro F1 of 0.564. This configuration generalized better than emotion-specific MLPs, likely due to capturing inter-class correlations. Key challenges included dataset label imbalance (e.g., dominance of "Fear") and performance drops between validation and test sets. Future work should explore synthetic data generation for minority emotions and newer BERT variants. This approach advances multi-hot emotion detection, with applications in opinion analysis and targeted sentiment modeling.

8 Ethical Considerations

Our investigation focuses solely on English text, and there may be bias for the contexts and emotions shown in the training data. The dataset may not be representative of all populations, with potential

biases in emotion detection for underrepresented groups. When using the model, it is important to consider the data used is anonymised and handled in compliance with privacy regulations. Emotion detection also has the potential for misuse for surveillance or manipulation. Further steps should be taken to prevent any biases identified during the evaluation process.

References

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Shivani Kumar, Md. Shad Akhtar, Erik Cambria, and Tanmoy Chakraborty. 2024. SemEval 2024 - task 10: Emotion discovery and reasoning its flip in conversation (EDiReF). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1933–1946, Mexico City, Mexico. Association for Computational Linguistics.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*, volume 5 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *arXiv preprint arXiv:2502.11926*.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Bhadresh Savani. 2021. Bert base uncased for emotion recognition. `https://huggingface.co/bhadresh-savani/bert-base-uncased-emotion`.

Lieve Van Woensel and Nissy Nevil. 2019. What if your emotions were tracked to spy on you? Technical Report PE 634.415, European Parliamentary Research Service.

Jane Wakefield. 2021. Ai emotion-detection software tested on uyghurs. BBC.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210. Published: February 28, 2006; Issue Date: May 2005.