Creating a Dataset for Multilingual Fine-grained Emotion-detection Using Gamification-based Annotation

Emily Öhman **Kaisla Kajava** **Jörg Tiedemann** **Timo Honkela**

University of Helsinki
firstname.lastname@helsinki.fi

Abstract

This paper introduces a gamified framework for fine-grained sentiment analysis and emotion detection. We present a flexible tool, *Sentimentator*, that can be used for efficient annotation based on crowd sourcing and a self-perpetuating gold standard. We also present a novel dataset with multi-dimensional annotations of emotions and sentiments in movie subtitles that enables research on sentiment preservation across languages and the creation of robust multilingual emotion detection tools. The tools and datasets are public and open-source and can easily be extended and applied for various purposes.

1 Introduction

Sentiment analysis and emotion detection is a crucial component in many practical applications but also defines a great challenge in natural language processing and artificial intelligence. Detecting emotions is crucial in human-computer interaction, and human behavior in communication is to a large degree affected by the emotional states that are created in a message. These states are typically fine-grained and fuzzy covering various dimensions of human feelings and attitudes. Nevertheless, it is often the practice to consider sentiments and emotions as very coarse and discrete features that can be detected with simple classifiers on a small scale of a few classes.

In our work, we focus on a high-dimensional model of emotions that allows a more natural and fine-grained classification and, furthermore, we tackle emotion detection in a multilingual setting. One of the biggest issues that stand in the way of creating reliable emotion detection algorithms is the lack of properly annotated datasets for training and testing purposes, especially in the case of the dimensionality that we consider and the multilingual support that we envision. This is the reason why we created *Sentimentator*, a new annotation tool that facilitates the efficient creation of appropriate datasets (Öhman and Kajava, 2018).

The main contribution of the paper is the framework based on a gamified environment that we develop to efficiently build large-scale resources. Our setup results in a *self-perpetuating* gold standard, which is initialized by seed sentences that are annotated by experts and augmented by crowd annotators. A combination of correlation-based scoring and ranking makes it possible to build datasets with weighted judgments based on the annotator confidence that we measure. Initial rankings are based on the comparison to seed annotation only but they will be adjusted dynamically once the correlation between crowd annotators allows to estimate further reliability scores. The main idea is that we can trust annotators that provide identical or at least similar judgments as other reliable annotators. With this scheme, we can move away from the use of limited seed sentences for confidence estimation to a more dynamic and self-perpetuating gold standard.

Another fundamental decision in our setup is the use of multilingual material on which to base our annotations. We are interested in the cross-lingual use of emotions and the development of multilingual classifiers (see Öhman et al., 2016). Therefore, we start with sentences extracted from movie subtitles (English originals in our case) for which we also have plenty of translations into a large number of languages. Movies contain a lot of emotional content and, as a side effect, it is interesting to see how that is reflected in subtitles and their translations.

Before presenting *Sentimentator* itself, we will first discuss related work and the theoretical framework we work with. The presentation of the seed/pilot dataset and its application for emotion detection follows the description of the tool and

Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 24–30

https://doi.org/10.18653/v1/P17

ends with a concluding discussion.

2 Related Work

Sentiment analysis is a widely studied task in natural language processing. Most of the existing datasets applied in sentiment analysis use binary or ternary annotation schemes (positive-negative, or positive, negative, and neutral) (Andreevskaia and Bergler, 2007), or some kind of combination of these (i.e. the addition of e.g. "mixed" (Saif et al., 2013)). This is not enough if the aim is to detect emotions rather than overarching sentiment (de Albornoz et al., 2012; Li and Hovy, 2017; Cambria et al., 2013). Furthermore, many of the existing datasets or tools (Munezero et al., 2015; Eryigit et al., 2013; Musat et al., 2012; Kakkonen and Kakkonen, 2011; Calefato et al., 2017; Saif et al., 2013; Abdul-Mageed and Ungar, 2017) are domain-dependent (often Twitter data) and/or document-level. Very few of these are also open data or open source.

An important questions is whether to show wider context or not. Boland et al. (2013) show that context can lead to the effect of double weighting for fine-grained annotations. For that reason, we also opted for the annotation of isolated sentences even though our tools would easily support other setups.

2.1 Crowd-sourcing Annotations

The annotation of datasets can be very costly and time consuming (Andreevskaia and Bergler, 2007; Devitt and Ahmad, 2008) if done by expert annotators. Crowd-sourcing can often be a cheaper alternative to hiring expert annotators, and has been used successfully by several researchers to create different types of datasets (Turney, 2002; Greenhill et al., 2014; Mohammad and Turney, 2013).

However, one issue with using non-experts to solicit annotations is that there is a risk of the quality suffering. Our solution to annotation-reliability related issues is gamification, which will be discussed in detail in section 3.

2.2 Theory of Emotion

The underlying theory of emotion for *Sentimentator* is Plutchik's theory of emotion (Plutchik, 1980). The eight core emotions he proposes are *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, and *trust*. He uses a wheel, or flower, to illustrate these emotions. For a more intuitive interface, we have inverted the wheel (see figure 1).

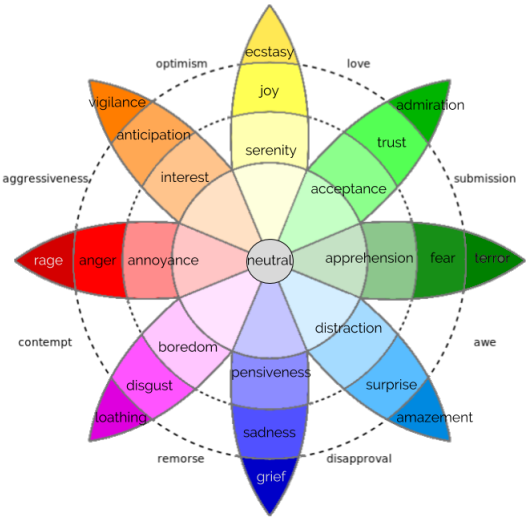

Figure 1: Inverted wheel of emotions

Although *Sentimentator* takes into account the intensity of the emotions, the complexity of the annotation task does not increase linearly with the number of classes this produces. It is possible to annotate for all 24+1 emotions, but only use the eight core emotions for classification, which was done very successfully by Abdul-Mageed and Ungar (2017).

Using Euclidean distance on the inverted wheel to calculate the similarity of annotations, we can see that annotations for *neutral* and low intensity emotions are in fact quite similar. This means we can avoid unnecessarily dismissing an annotation as noise, as might be the case if a more traditional interface was used where *neutral* was a separate category from low-intensity emotions.

2.3 Classification

Table 1 shows the accuracies achieved by a few other multidimensional approaches (generally those of Ekman (1971)) using various classification methods such as SVMs, neural networks, maximum entropy, Naïve Bayes, and k-Nearest Neighbor to name a few. When more classes are included in a model, accuracies achieved are typically lower than what binary or ternary models of sentiment analysis achieve (Purver and Battersby, 2012; Tokuhisa et al., 2008), with the exception of Abdul-Mageed and Ungar (2017) who apply a gated recurrent neural network model.

[1]Quoted in Purver and Battersby (2012)

Study	classes	accuracy
Go et al. (2009)	2	82.2-83%
Danisman and Alpkocak (2008)	6	32%
Chuang and Wu (2004)	6+1	56-74%
Chuang and Wu (2004)	6+1, audio	81.5%
Ansari 2010[1]	6	81%
Purver and Battersby (2012)	6	varies
Seol et al. (2008)	8	45-65%
Abdul-Mageed and Ungar (2017)	8	95.68%
Tokuhisa et al. (2008)	10	up to 80%
Abdul-Mageed and Ungar (2017)	24	87.58%

Table 1: Accuracies achieved by previous studies

Although the data and methods are different, it seems reasonable to expect accuracies between 30-70% depending on the category for an initial multiclass classification. Depending on the availability, we will try to apply our model to the same datasets that have been used in the studies listed in table 1 in order to directly compare results.

3 Gamifying the Annotation Platform

Our goal is to implement efficient crowd-sourcing through *gamification*. Gamification refers the use of game elements in environments that are not typically games (Deterding et al., 2011; Hamari and Koivisto, 2013). Previous research shows that one can achieve a high number of quality annotations by non-experts by using carefully considered gamified aspects such as (1) Relatedness (connected to other players), (2) Competence (mastering the game problems), and (3) Autonomy (control of own life) (Musat et al., 2012).

Robson et al. (2015) posits that gamification can change behavior by tapping into motivational drivers of human behavior: reinforcements and emotions. The emotions we want to elicit are of course enjoyment, but negative emotions such as disappointment can also increase commitment and a desire to increase one's competence. Similarly, both positive and negative reinforcements increase repetitive behavior in players (Robson et al., 2015).

Our platform offers players leaderboards and statistics about both their immediate and longterm progress (relatedness). Progress, rank and prestige are important measures that help players feel compensated for the work they are doing within the game (competence). Rank has an additional function in our platform; as we lack a gold standard against which to compare the annotations we receive, we use rank to determine noisy annotations and noisy annotators. Furthermore, the player can see how each of their choice affects their standing in the ranks.

The dataset that the experiments in this paper rely on is our validated seed sentences. These sentences will be used as a type of seeded gold standard. What this means is that annotators will annotate both non-seed sentences and seed sentences. They receive a score from their annotation based on similarity with gold annotations that determines their rank. In practice, rank is equivalent with confidence level. With enough participants, players will also be ranked according to how closely their annotations match those of high-ranked annotators. We, thus, have the option to include only the least noisy, highest quality annotations in our dataset.

In order to compare annotations, we map them on the inverted Plutchik's wheel we propose in Figure 1 projected on a standard two-dimensional space with coordinates $[x, y]$, where the least intense emotions are at the center and the most intense at the tips of the petals. The origin of our emotion space is located in the center of the wheel and represents the case of a neutral expression.

We can then calculate the distance $D(G_x, G_y)$ between any pair of points corresponding to emotion labels G_X and G_y by computing the Euclidean distance normalized by the maximum distance that can be observed between opposite emotions with maximum intensity. Assuming that geometric location expresses the relatedness of emotions, this distance metric takes into account all different types of similarities/dissimilarities between annotations, including labels that combine neighboring emotions.

The distance metric is the basis for the computation of annotation confidence C_x for new annotation G_x that we obtain. We define annotation confidence as

$$C_x = R_{annotator} * \frac{1}{N} \sum_{n=1}^{n} (1 - C_n * D(G_x, G_n))$$

where $G_n \in \{G_1, .., G_N\}$ are annotations of the same instance (sentence) from the current gold standard with corresponding confidence scores $\{C_1, .., C_N\}$.[2] In other words, we add an averaged penalty for annotations that differ from existing gold annotations weighted by their confidence scores. Note that our seed annotations G_s obtain

[2]Note that we set $C_x = R_{annotator}$ if $N = 0$.

a perfect confidence $C_s = 1$. Another component of the confidence score is the rank of the annotator based on the score $R_{annotator}$. This score is initialized with one and will be updated by each submitted annotation. Currently, we use a simple average over annotation confidence scores of that particular annotator:

$$R_{annotator} = \frac{1}{M} \sum_{m=1}^{M} C_m$$

Our self-perpetuating gold standard with ranking-based confidence ratings will reduce the need for manual screening and will ensure that we can receive consistent emotion annotations with a measurable confidence attached to them.

4 Creating the Dataset

For the seed data, we used the following procedure: On completed expert annotation, another expert annotates the same sentences with the data order randomized. Ambiguous sentences were reviewed and the correct class was agreed upon. In some cases where no agreement could be reached the sentence was excluded from the seed dataset.

Our data collection will be unique in that it will provide a fine-grained multi-dimensional open source dataset for sentiment analysis and emotion detection in various languages. Annotation is on-going and the first real dataset will be available later in 2018. For now we have a set of sentences with validated annotations that we will use as our seed data to get the gamified annotation started. This dataset has already been used to investigate sentiment preservation in Finnish, French, and Italian (Kajava, 2018).

We wanted to make the dataset as useful as possible to as many researchers as possible from the beginning. This is why we selected an open parallel corpus, namely the OPUS movie subtitles corpus (Tiedemann, 2012; Lison and Tiedemann, 2016). From this collection, approximately 9,000 English sentences were annotated into the following emotion classes: *anger*, *anticipation*, *disgust*, *joy*, *fear*, *sadness*, *surprise*, and *trust*. This yielded a preliminary dataset of between 649 to 908 sentences per class (see table 4). For the classification experiments, we did not take into account the measure of emotion intensity that we introduce later in our annotation framework, which is also part of the seed sentence annotation.

	ang	ant	dis	fea	joy	sad	sur	tru	Total
train	816	739	775	615	876	633	583	737	5,774
test	92	84	87	70	99	72	66	83	653

Table 2: Emotion distribution in seed data

We also keep the metadata and therefore all sentences can be paired with a particular movie, genre, time period, as well as its counterpart in another language. This is valuable information for future research avenues.

We expect to have a full dataset by the end of the year as we will be collecting at least 100 000 annotations in September-October of 2018 from crowd annotators (students). Snow et al. (2008) suggest using four non-experts to match the quality of one expert annotator, however, gamification, seed sentences, and rank-validation means that fewer annotations per sentence might be sufficient using our platform. Inter-annotator agreement is on average around 70-90% depending on the type of annotation and who is doing the annotation work (expert vs. non-expert) (Nowak and Rüger, 2010), and this is where annotator agreement was for our data as well, varying between classes.

Based on initial timed annotations, a typical annotator can be expected to annotate up to 10 sentences per minute. This means that it only takes just over one hour to annotate around 600 sentences. If every annotator is asked to annotate 1000 sentences, this should not take more than a few hours each on average taking the learning curve into account. The students are encouraged to annotate in languages other than English as well, resulting in at least two or three separate datasets with an expected minimum of 40 000 annotations each. We currently have preliminary datasets for English, Italian, French, and Finnish.

5 Validation of the Data Quality

In order to test the quality of the data for the purpose of developing an automatic emotion detector we ran some initial experiments using our annotated seed data for training, and evaluating standard multi-class classifiers. The data was tokenized and lowercased as a preprocessing step. We selected Multinomial Naïve Bayes (NB) and Multilayer Perceptron (MLP) classifiers for our experiments. For the classifiers we use the scikit-learn (Pedregosa et al., 2011) machine learning toolkit. In both classification scenarios, the data was split into class-stratified training and test sets of 90%

and 10%, respectively.

The MLP network used in this work is a three-layer network. The model creates a lexicon from the dataset using a bag-of-words approach, employing it for extracting a set of features for each class. We use Adam for training the network and apply Rectified Linear Unit (ReLu) activiation functions in the hidden layers of feed-forward network.

Classifier	Accuracy
NB	0.5069
MLP	0.5023

Table 3: Overall classification accuracy

As can be seen in table 3, the baseline classifiers perform reasonably well for such a small data set and such a fine-grained task. Note that we did spend any time on optimizing features and hyperparameters to obtain a better performance. The purpose of this study is entirely to test the feasibility of fine-grained classification and validity of our seed data.

In the confusion matrix for the best performing classifier (see Table 4), we can see that there is some significant confusion between anger and fear, between disgust and sadness and also surprisingly between trust and fear. These are the same classes that others have struggled to distinguish (e.g. (Purver and Battersby, 2012)), which is reassuring that our data is in good shape. With this promising performance on our pilot data set (despite its limited size) we are encouraged to proceed with future experiments and the more fine-grained distinctions we propose that take intensity into account.

These experiments also demonstrate that the seed data is sufficient for initial classifications and that we can go ahead in developing our gamified strategy of getting more annotations based on correlations between annotators and their level of trust, which will initially be based on the comparison to the validated seed sentences.

6 Conclusions

In this paper, we present an open annotation tool for fine-grained emotion detection and a dataset of seed sentences that can be used to gamify the annotation efforts. The classification results show that the dataset is reliable enough to be used as seed sentences, indicating that gamification based on the seed data is a viable option for compiling sentiment datasets, and that multidimensional classification yields acceptable results even with a small dataset. However, already with our limited validation experiments we can see that the choice of model and learning algorithm influences the quality of the resulting classifier. This is an important outcome that needs to be considered when designing tools with scarce resources. We will continue to monitor performance to measure the impact of gamification and cross-lingual transfer on classification performance.

ang	**ant**	**dis**	**fea**	**joy**	**sad**	**sur**	**tru**	<- classified as
62	3	10	3	5	1	3	5	**anger**
9	**43**	6	1	5	1	5	14	**anticipation**
25	6	**29**	5	9	3	3	7	**disgust**
11	8	6	**20**	2	6	5	12	**fear**
2	4	5	1	**77**	2	3	5	**joy**
7	2	8	5	12	**30**	2	6	**sadness**
5	4	10	4	13	1	**25**	4	**surprise**
5	3	5	3	16	3	2	**45**	**trust**

Table 4: Confusion matrix for NB-based classification of the test set.

7 Discussion and Future Work

As our system collects both coarse (ternary) sentiment annotations, and fine-grained emotion annotations, a future option could be to apply the successful approach demonstrated by Tokuhisa et al. (2008), and utilize the coarser sentiment polarity to pre-classify data before emotion classification, and then implement a k-nearest neighbors algorithm on the larger dataset. Our platform does not show context by default, however, it is easy to add the context of the target segment to be annotated if required for a different type of project.

It might be valuable to re-annotate parts of the data showing additional context to check the impact on annotation and annotator confidence. Tokuhisa et al. (2008) found that in their data, context-dependent samples were useful for training their classifier and yielded slightly higher accuracies than their non-context-dependent data. Finally, we will also consider models that make use of sentence-internal relations to improve the classification results. In particular, we will investigate the use of sequence models and gated recurrent networks as proposed by Abdul-Mageed and Ungar (2017).

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 718–728.

Jorge Carrillo de Albornoz, Laura Plaza, and Pablo Gervás. 2012. Sentisense: An easily scalable concept-based affective lexicon for sentiment analysis.

Alina Andreevskaia and Sabine Bergler. 2007. Clac and clac-nb: Knowledge-based and corpus-based approaches to sentiment tagging. In *Proceedings of the 4th international workshop on semantic evaluations*, pages 117–120. Association for Computational Linguistics.

Katarina Boland, Andias Wira-Alam, and Reinhard Messerschmidt. 2013. Creating an annotated corpus for sentiment analysis of german product reviews.

Fabio Calefato, Filippo Lanubile, and Nicole Novielli. 2017. Emotxt: a toolkit for emotion recognition from text. *arXiv preprint arXiv:1708.03892*.

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.

Ze-Jing Chuang and Chung-Hsien Wu. 2004. Multimodal emotion recognition from speech and text. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 9, Number 2, August 2004: Special Issue on New Trends of Speech and Language Processing*, 9(2):45–62.

Taner Danisman and Adil Alpkocak. 2008. Feeler: Emotion classification of text using vector space model. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, volume 1, page 53.

Sebastian Deterding, Miguel Sicart, Lennart Nacke, Kenton O'Hara, and Dan Dixon. 2011. Gamification. using game-design elements in non-gaming contexts. In *CHI'11 extended abstracts on human factors in computing systems*, pages 2425–2428. ACM.

Ann Devitt and Khurshid Ahmad. 2008. Sentiment analysis and the use of extrinsic datasets in evaluation. In *LREC*.

Paul Ekman. 1971. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press.

Gülsen Eryigit, Fatih Samet Cetin, Meltem Yanik, Tanel Temel, and Ilyas Çiçekli. 2013. Turksent: A sentiment annotation tool for social media. In *LAW@ ACL*, pages 131–134.

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

Anita Greenhill, Kate Holmes, Chris Lintott, Brooke Simmons, Karen Masters, Joe Cox, and Gary Graham. 2014. Playing with science: Gamised aspects of gamification found on the online citizen science project-zooniverse. In *GAMEON'2014*. EUROSIS.

Juho Hamari and Jonna Koivisto. 2013. Social motivations to use gamification: An empirical study of gamifying exercise. In *ECIS*, page 105.

Kaisla Kajava. 2018. Cross-lingual sentiment preservation in binary and multi-dimensional classification.

Tuomo Kakkonen and Gordana Galić Kakkonen. 2011. Sentiprofiler: creating comparable visual profiles of sentimental content in texts. *Language Technologies for Digital Humanities and Cultural Heritage*, 62:189–204.

Jiwei Li and Eduard Hovy. 2017. Reflections on sentiment/opinion analysis. In *A Practical Guide to Sentiment Analysis*, pages 41–59. Springer.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

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

Myriam Munezero, Calkin Suero Montero, Maxim Mozgovoy, and Erkki Sutinen. 2015. Emotwitter–a fine-grained visualization system for identifying enduring sentiments in tweets. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 78–91. Springer.

Claudiu-Cristian Musat, Alireza Ghasemi, and Boi Faltings. 2012. Sentiment analysis using a novel human computation game. In *Proceedings of the 3rd Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources and Their Applications to NLP*, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566. ACM.

Emily Öhman, Timo Honkela, and Jörg Tiedemann. 2016. The challenges of multi-dimensional sentiment analysis across languages. *PEOPLES 2016*, page 138.

Emily Öhman and Kaisla Kajava. 2018. Sentimentator: Gamifying fine-grained sentiment annotation. In *Digital Humanities in the Nordic Countries 2018*. CEUR Workshop Proceedings.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31.

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491. Association for Computational Linguistics.

Karen Robson, Kirk Plangger, Jan H. Kietzmann, Ian McCarthy, and Leyland Pitt. 2015. Is it all a game? understanding the principles of gamification. *Business Horizons*, 58(4):411 – 420.

Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. 2013. Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold.

Yong-Soo Seol, Dong-Joo Kim, and Han-Woo Kim. 2008. Emotion recognition from text using knowledge-based ann. In *ITC-CSCC: International Technical Conference on Circuits Systems, Computers and Communications*, pages 1569–1572.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.

Jorg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. 2008. Emotion classification using massive examples extracted from the web. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 881–888. Association for Computational Linguistics.

Peter D. Turney. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.