Proceedings of the LREC 2020 Workshop "Citizen Linguistics in Language Resource Development", pages 15–25
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Objective Assessment of Subjective Tasks in Crowdsourcing Applications

Giannis Haralabopoulos, Myron Tsikandilakis, Mercedes Torres Torres, Derek McAuley
University of Nottingham
name.surname@nottingham.ac.uk

Abstract

Labelling, or annotation, is the process by which we assign labels to an item with regards to a task. In some Artificial Intelligence problems, such as Computer Vision tasks, the goal is to obtain objective labels. However, in problems such as text and sentiment analysis, subjective labelling is often required. More so when the sentiment analysis deals with actual emotions instead of polarity (positive/negative) . Scientists employ human experts to create these labels, but it is costly and time consuming. Crowdsourcing enables researchers to utilise non-expert knowledge for scientific tasks. From image analysis to semantic annotation, interested researchers can gather a large sample of answers via crowdsourcing platforms in a timely manner. However, non-expert contributions often need to be thoroughly assessed, particularly so when a task is subjective. Researchers have traditionally used 'Gold Standard', 'Thresholding' and 'Majority Voting' as methods to filter non-expert contributions. We argue that these methods are unsuitable for subjective tasks, such as lexicon acquisition and sentiment analysis. We discuss subjectivity in human centered tasks and present a filtering method that defines quality contributors, based on a set of objectively infused terms in a lexicon acquisition task. We evaluate our method against an established lexicon, the diversity of emotions - i.e. subjectivity- and the exclusion of contributions. Our proposed objective evaluation method can be used to assess contributors in subjective tasks that will provide domain agnostic, quality results, with at least 7% improvement over traditional methods.

Keywords: Natural Language Processing, Crowdsourcing, Lexicon, Subjectivity, Objectivity

1. Introduction

Data is the most sought-after commodity of the digital era. Through interaction, expression and reasoning we produce varying types of data. From a philosophical standpoint, there are two main categories of information embedded in data: objective and subjective information. Objective information relates to empirical facts and their measurement, while subjective information relates to the personal experience and expression of thoughts, opinions and emotions. In the digital space, the objectivity and subjectivity of the information can be linked to human factors. As humans interact with the digital world, the information they share is subject to analysis from scientists and commercial stakeholders. The most common analysis performed, in human submitted digital information, is sentiment analysis (Yue et al., 2018).

Sentiment analysis aims to explore the subjective emotions conveyed in information (Chaturvedi et al., 2018; Yoshino et al., 2018), such as multimedia or simple text sources (Miao et al., 2018; Öztürk and Ayvaz, 2018). With regard to textual information, crowdsourcing is most frequently used to obtain the emotion conveyed in paragraphs of text (Li et al., 2018). Their analysis requires the emotional labelling of full sentences, part of sentences, or terms (Hazarika et al., 2018).

If labelling within the corpus is extensive, then supervised sentiment analysis methods can be applied (Zhao et al., 2018). On the other hand, if no labelling is available, unsupervised methods will need to be employed (Fernández-Gavilanes et al., 2018). If the labelling required to annotate the corpus is extensive, then an unsupervised approach might be a better method (Fernández-Gavilanes et al., 2018). However supervised learning generally obtains better results in most machine learning problems (Schouten et al., 2018).

Expert labelling is both expensive and time consuming (Palan and Schitter, 2018). As an alternative, crowdsourcing enables scientists to recruit a higher number of individuals to improve the quality of the labelling process through redundancy. Crowdsourcing is the process of non-expert annotators contributing to scientific tasks (Howe, 2006). Crowdsourcing platforms provide access to a diverse range of contributors (Peer et al., 2017). Data gathered for sentiment analysis favors distinct classes rather than a distribution of classes (Koltsova et al., 2016; O'Leary, 2016). Even when the requested data spans through several categories, the results are filtered based on a gold standard (Tang et al., 2015; Maynard and Bontcheva, 2016).

Polarity, i.e. positive and negative emotion, is a common topic of interest that leads to refined polarity and extended to pure emotion or beyond polarity analysis (Basile et al., 2018; Sharma and Chakraverty, 2018). In polarity-based annotation tasks, contributors are tasked with deciding between a positive or a negative label (Budhi et al., 2018), Conversely, in a refined or pure emotion analysis annotators are labeling text using either a scale from negative to positive, or the provided emotional list respectively (Ghosal et al., 2018).

The gold standard is used to filter spam or dishonest responses. It is based on predefined expected answers. It is widely used in image analysis and crowdsourcing applications (Ghosh et al., 2015). It has also been used in the subjective evaluation of emotional information (Calefato et al., 2017), alongside with majority voting (Zamil et al., 2019), to determine the most appropriate label for a term, group or sentence. Majority voting methods appoint the most annotated emotion as the corresponding emotion label. Information loss occurs in both methods since the annotations that are not part of the major/gold class are excluded. Additionally, these methods fail to address the subjective nature of emotion labelling.

We argue that the aforementioned dominant class selection

methods disregard human subjectivity. In a subjective labelling task, single class or ground truth do not accurately portray the diversity of human evaluation. We propose the use of emotion vectors to retain subjectivity, and the evaluation of contributions based on infused objectively emotional terms. We perform a set of subjective crowdsourcing tasks to assess our proposed method, in which we evaluate participants through their performance solely on terms of objective emotional significance.

The main contributions of this paper are: a contributor evaluation method for subjective crowdsourcing tasks and the use of objective terms based on the subjective task itself. We also highlight the differences of our quality assessed resource when compared to an established pure emotion lexicon.

2. Subjectivity

Subjectivity has been defined as "[...] the lived diversity in experience due to the physical, political and cultural context of [an] experience" (Ellis and Flaherty, 1992). This definition could be a rally point for enabling us to understand the concept of emotion as a universal experience with subjective variability.

For example, there are widely accepted concepts of "universals" in research relating to emotion. These include the theory of universal emotions proposed by Ekman and Friesen (Ekman and Friesen, 1971) and the theory of primary bipolar emotions as suggested by Plutchik (Plutchik, 1980). According to these seminal social and psychological theories anger, fear, happiness (or joy), disgust, sadness and surprise, and also trust and anticipation are emotions that can be encountered cross-culturally (Ekman and Keltner, 1997). These emotions are also suggested to have shared evolutionary neural and physiological functions. These functions involve automatic and involuntary responses to danger (fear) and sudden environmental changes (surprise), social communication of positive (happiness, joy, trust) and negative states (anger, sadness) and responses to potentially harmful pathogens and nourishment (disgust) (Pessoa and Adolphs, 2010). In a sense these emotions are a "universal language".

The aforementioned definition of subjectivity included the phrase "cultural diversity". Cultural diversity is one of the most widely studied correlates of subjectivity for emotional annotation (Elfenbein, 2017). Contemporary research has found that although there are basic and/or primary emotions that could, indeed, be a "universal language", there are also culture-specific "dialects". These dialects are used for displaying these emotions in terms of facial expressions (Elfenbein and Ambady, 2002). They are also used for communicating culturally-appropriate emotional intensity in written and verbal expressions (Elfenbein and Luckman, 2016). These cultural dialects are suggested to confer an own-culture emotional recognition advantage in response to own-culture stimuli. They are also, arguably, suggested to confer an other-culture emotional recognition bias in response to other-culture stimuli that are distinctly different to the culture of the respondent (Keith, 2019). This is suggested to occur due to the non-convergent social evolution that takes place in different geographical areas. This could mean that although we all understand basic emotions such as fear and happiness, we may display (show) and decode (understand) these emotions differently due to our cultural background (Elfenbein, 2017).

For example, previous research has shown that Western individuals use high-intensity emotional words during social interactions (Semnani-Azad and Adair, 2013). It has also been suggested that Western individuals are not likely to recognise low-intensity expressions of emotion; possibly because these are not accurately discriminated as communicating salient emotional information (Knapp et al., 2013). Conversely, previous research has shown that Eastern individuals use context-specific positive emotional expressions in their social interactions (Masuda et al., 2008). It has also been suggested that Eastern individuals are not likely to acknowledge that a negative in valence expression was part of a social interaction. This is suggested to occur because the acknowledgement would necessitate a negative and culturally inappropriate social response (Matsumoto et al., 2013).

In the same manner, the valence and the meaning we attribute to words and images can be different between cultures (Lauka et al., 2018), between genders (Chaplin, 2015) and between age groups (Silvers et al., 2016). For example, the word "fight", as well as images that show virtual violence (Yao et al., 2017), are often considered to convey positive high arousal in young male respondents. The same stimuli have been shown to elicit neutral and negative emotional responses in older adults, irrespective of gender, and female participants; irrespective of age (Gohier et al., 2013; Reidy et al., 2016). Similar effects, such as differential positive or negative or neutral responses to high-arousal words, have also been reported due to differences in political orientation, religious affiliation and emotional sensitivity (Smith, 2015).

Subjectivity can also occur in response to seemingly innocuous stimuli due to differences in physical experiences such as bodily needs and even illness (Teo, 2018). For example, the on-screen presentation of the, arguably, neutral words "dinner" and "food" has been shown to elicit idiosyncratic annotating, behavioural, physiological and neural responses in specific populations. Individuals who are suffering from an eating disorder (Canetti et al., 2002) and also healthy individuals who have been subjected to mild food deprivation and transient insulin-induced hypoglycemia (Brody et al., 2004) have been shown to label the words "dinner" and "food" as high emotional intensity items.

Accordingly, subjectivity is an important, multi-sided and possibly unavoidable aspect of human interactions. The challenge at hand is how to best incorporate subjectivity in our coding-response framework without treating it as participant error or response bias while at the same time controlling for participant error and response bias (Rouder et al., 2016).

3. Sentiment Analysis

Sentiment analysis is, in its core, a subjective process (Mihalcea et al., 2007). As mentioned above, sentiment analysis can be performed with or without manual labelling; such as supervised or unsupervised methods. Supervised

sentiment analysis and other similar methods that utilise a lexicon require a level of manual input. That manual input can be obtained by the scientists themselves, or via crowdsourcing. Crowdsourcing has been used as a method to obtain a large number of manual inputs from an equally large number of contributors. Multiple contributors can be used to obtain an emotion per word association (Kiritchenko and Mohammad, 2017), and a ranked order of words on a best to worst emotional scale. Crowd contributors can identify events, perform predictions and provide emotional annotations for the available data (Schumaker et al., 2016). Subjective topics, such as the discussion and promotion of creative ideas, can also be analysed via the crowd (O'Leary, 2016).

Often, the crowdsourcing inputs need to be evaluated, particularly when the task is objective. The gold standard method described in the introduction is one form of manual evaluation. The evaluation is usually performed by individuals with certain expertise in the task. The definition of experts is most commonly vague and their appointment is often biased. For example, previous publications have provided such definitions of expertise as "three experts in the smartphone industry" (Chamlertwat et al., 2012), "the two authors plus one other colleague" (Diakopoulos and Shamma, 2010), "10 financial experts" (Ranco et al., 2015), "post-graduate students who have at least three years' experience for the respective product domains" (Lau et al., 2014), or did not include further elaboration in regard to the description of the included experts (Kang and Park, 2014; Prabowo and Thelwall, 2009; Hutto and Gilbert, 2014; Caselli et al., 2016).

Expert evaluation of subjective tasks should be reconsidered (Eickhoff, 2018). The relevance (Luhrmann, 2006) and role (Kittur et al., 2008) of expert assessment in subjective topics, such as sentiment analysis, is debated (for a comprehensive review, see Hetmanck's review (Hetmank, 2013)). The exact relationship between the experts and the authors, and the prevalent implicit bias of collaborative relations often remain undisclosed. In the case that the experts are not affiliated with the authors but are externally hired (Haralabopoulos et al., 2018; Haralabopoulos and Simperl, 2017) implicit bias could occur due to the monetary reward involved.

4. Proposed methodology

We propose the evaluation of crowd contributors on a set of objective terms. The objective terms can be the emotions themselves or they can stem from the emotion itself, e.g. "joyous" from "joy", "angry" from "anger". A random number of terms is injected into a simple emotion annotation task hosted in Amazon Mechanical Turk[1]. The objective terms appear randomly during the task, are always followed by a subjective term and rotate over emotions, Table 1.

[1] https://www.mturk.com/

Emotion	Objectively Emotional Terms
anticipation	anticipate anticipating anticipated
joy	joyful joyous joy
trust	trusted trustees trusting
fear	feared fears fearful
sad	sad sadly saddened
disgust	disgusted disgusting disgustful
anger	angered angering angerful
surprise	surprised surprising surprisingly

Table 1: Objective Terms

To identify the optimal number of injected terms, we perform four distinct tasks with varying levels of objective terms injected. We ask contributors "What emotion better describes the current word?". The allowed answers are the eight basic emotions, as defined by Plutchik (Plutchik, 1980). We refrained from including a neutral emotional state because it has been shown that there is a low neutrality consensus for text (Valdivia et al., 2018). We assess each contributor with three different methods, majority voting, threshold, and one objective evaluation process.
Let W be a worker with $\{a_1, a_2, ..., a_j\}$ annotations $a \in \{1, 2, ..., k\}$ and $k \in \mathbb{Z}$, towards a set of terms $T = \{t_1, t_2, ..., t_j\}$. Each method is formulated as follows:

4.1. Majority Voting

For each term t the majority class t_m is defined by:

$$t_m = r_t \text{ with } r \in \{1, 2, ..., k\} \tag{1}$$

$$P_t(r_t) \geq P_t(a_n) \forall\ a_n \in \{a_1, a_2, ..., a_j\}\ \&\ a_n \neq r_t, \tag{2}$$

where $P_t(x)$ is the probability of class x appearing in the annotations of term t.
Majority voting discards answers and contributors that were not in agreement with the majority of annotations. Each worker is assessed based on the majority classes that were in line with the supplied annotations; e.g. a task requester can discard annotations from users that disagreed with the majority classes at a given percentage. Most frequently, the majority class is also defined as the "correct" class for each term.

4.2. Threshold

Let $h \in [0, 1]$ be a predefined threshold. A worker W has their annotations discarded if in:

$$\{a_1, a_2, ..., a_j\}\ \exists\ a_n \mid P(a_n) \geq h \tag{3}$$

Threshold filtering forces diversity, as requesters can discard contributors with a fixed percentage of annotations in a single answer.

4.3. Objective Annotator Evaluation

To apply an objective evaluation of annotators, we inject $\{t'_1, t'_2, ...\}$ terms, into T, that confine the emotional stimuli (Brosch et al., 2010). The classes l' of t' are predetermined, $\in 1, 2, ...k$, and we judge annotator performance via a micro-averaged F1 method:

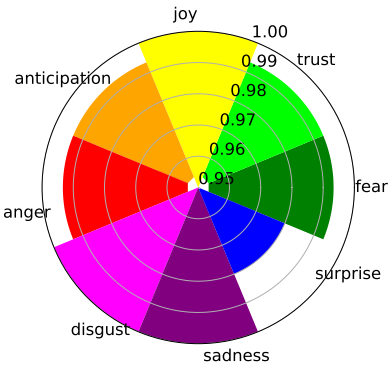


Figure 1: 25%

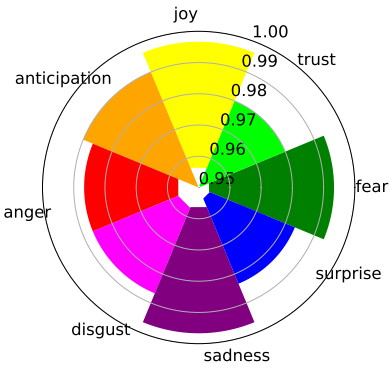


Figure 2: 33%

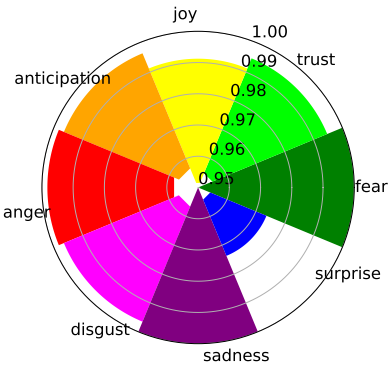


Figure 3: 50%

F1 scores for different objective term injection ratio

		Class	
		l'	$\neq l'$
Annotated	$a' = l'$	*TP*	*FP*
	$a' \neq l'$	*FN*	*TN*

4.3.1. F1 Score

$$precision = \frac{TP}{TP + FP} \tag{4}$$

$$recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{6}$$

The algorithmic process can be seen in Algorithm 1. We have a crowdsourcing task, performed by a number of participants. The evaluation method, can be one of the three mentioned above, aims to identify honest contributors. Each participant is evaluated and if deemed honest, is added to the set of quality contributors. Their answers are then returned to the requesters. I.e. the objective terms inside the task function as an honesty assessment.

Algorithm 1: Selection Process Pseudo-Algorithm

Task() = Crowdsourcing Task;
Eval() = Evaluation method;
QC = Set of Quality Contributors;
for *participant* ***in*** *Task()* **do**
  Eval(participant);
  if *Eval(participant)* ***is*** *True* **then**
    add participant **to** QC;
  end
end
return *Task(QC)*

5. Experiment

We inject a set of objective terms, Table 1, into a subjective dataset. The simplicity in task evaluation yields better results (Finnerty et al., 2013) and provides task consistency. Contrary to usual gold standard methods where generic questions are asked to assess the attention of contributors (Aker et al., 2012). The design of the task is based on left to right saccadic movements, consistent with the natural reading patterns of participants as reported in previous research (Starr and Rayner, 2001; White et al., 2015; Smith and Elias, 2018). Although we manually created the objective terms group and regardless of the domain or the task, we can easily obtain a set of objective terms based on the stems and suffixes of the answers.

We choose the subset of common terms found in emotion lexicons, NRC(Mohammad et al., 2013) and PEL (Haralabopoulos et al., 2018; Haralabopoulos and Simperl, 2017). Both lexicons are multi emotion labelled and enable us to select terms with the highest emotional variation, i.e. words with the most diverse emotions annotation.

We created four sub-datasets, based on the ratio of objective to subjective terms. One had no objective terms injected (0%), one had a quarter of subjective terms injected (25%), one had one objective term per two subjective terms (33%) and the final set had the same number of subjective and objective terms (50%). Each term received 10 annotations from 10 different contributors and maximum time per question was 120 seconds.

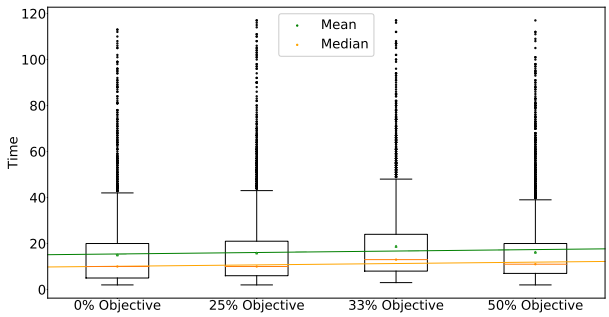


Figure 5: Time Required for Subjective Answers

We present an analysis of the annotators' performance followed by an evaluation section for the results. The evaluation is divided in three parts: a direct correlation analysis of the obtained results and NRC emotion vectors, an emo-

tional diversity analysis and finally a redundancy and exclusion analysis.

5.1. Contributors

The time required, per contributor, to answer each question was analogous to the ratio of injected terms, Figure 5. As the contributors encountered more objective terms, their mean answer time requirement - from 0% to 50% objective terms - went from 14.97s to 16.13s and the median response time from 10s to 11s. An increase of 10% across both metrics indicates an increase of contributors' attention to the task.

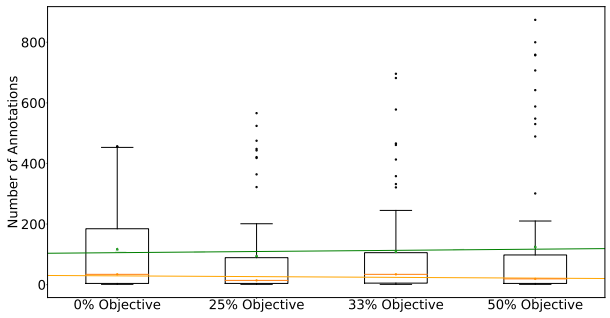


Figure 6: Number of Annotations per Contributor

Tasks occupied an analogous - to the injected terms - number of participants. The 0% task had 39 participants, 25% had 61, while 63 and 73 people contributed to 33% and 50% tasks respectively. Attention requirements of the task negatively affected participation. The task design and layout was consistent throughout all of the tasks, therefore no varying complexity or difficulty factor existed. Due to the increasing number of participants, as the number of injected term increased, the median number of contributions per participant decreased. The mean number of contributions is affected by a large number of major outliers, Figure 6.
With regard to the distribution of objective and subjective terms contributions per participant, the results follow the corresponding injection ratios, slightly affected by contributors with less than 20 subjective answers. Each contributor encountered a median of 20%, 30% and 50% objective terms for their respective injection ratios, Figure 7. The y-axis is the ratio of objective terms to total terms, as encountered by each participant.

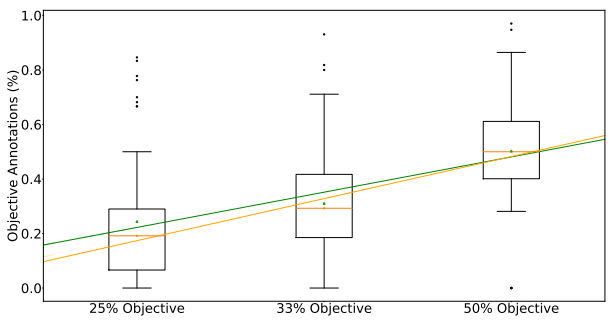


Figure 7: Percent of Objective Annotations per Contributor

The performance of contributors, as measured by our F1 score, was fairly consistent. On average the contributors managed to correctly annotate >96% of the objective terms across all emotions, Figures 1, 2 and 3. The F1 score for surprise-related objective annotations (Table 1) was low in all three different injection ratios. The objective terms for 'sadness', 'fear', and 'joy' had >99% F1. A small variation was observed on the annotation of objective trust terms, especially in the 33% ratio. The number of objective terms does not seem to affect the F1 scores monotonically, since the F1 scores for the objective terms of 33% were worse than those for 50% and 25%. The excluded participants based on a required perfect F1 score where 14 on the highest 50% objective ratio, 11 at 33% and 3 at the 25%.

Injection Ratio	Correct annotations(%)
50	0.9939%
33	0.9892%
25	0.9942%

Table 2: Correct annotation of objective terms for different injection ratios

The distribution of emotions was similar, irrespective of the injection ratio, Figure 12. However, when annotators encountered no objective terms in their task mostly annotated subjective terms as related to trust, joy and disgust. The highest injection ratio (50%) had lower trust and disgust annotation which were redistributed to anger, anticipation and fear. The ratio of objective terms didn't seem to affect the performance of contributors. The overall objective classification accuracy remained around and above 99%, Table 2.

5.2. NRC Correlation

We compare our results to the NRC lexicon (Mohammad et al., 2013). The Spearman's Rho correlation is calculated for each term vector in our results, against the same term vector in NRC. For example, the term '*absolution*' had the following emotional vector in one of our tasks: *[0.0, 0.2, 0.6, 0.0, 0.0, 0.1, 0.1, 0.0]*, and the following vector: *[0.0, 0.5, 0.5, 0.0, 0.0, 0.0, 0.0, 0.0]* in NRC, a correlation of 0.8109. We present Interquartile Range plots for all 456 term correlations in our results and a summarising table with mean and median per term correlation.

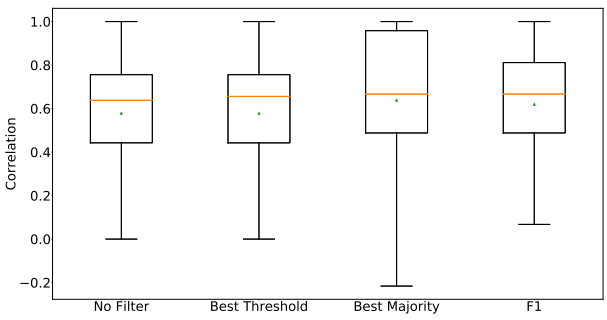


Figure 8: I.Q.R. of per term correlation for all filtering methods, 50% objective terms

For each task of the four crowdsourcing tasks of different injection ratios, we compare the performance of four differ-

	a		b	
Method	Mean	Median	Mean	Median
No filter	0.5781	0.6381	0.5781	0.6381
20% Threshold	0.5578	0.6547	0.5578	0.6547
100% Majority	**0.6498**	0.6547	0.0656	0.0660
F1	0.6191	**0.6667**	**0.6191**	**0.6667**

Table 3: Comparing Spearman's Rho (a) and Adjusted Score (b) for 50% injection ratio

ent filtering methods. No filter method refers to the results as received from the crowdsourcing task. The X% threshold entails the removal of all annotators that annotated more than X% of their terms with the same emotion. To determine the best threshold method for each injection ratio, we calculate the correlation for four different thresholds 20% - 30% - 40% - 50%. For each term, after the end of the task, we determine one or more major emotions. By comparing the annotations of each contributor in relation to the major class(es) of each term we acquire a per contributor majority agreement factor. To obtain the best majority method we calculate the correlation for 100% - 90% - 80% -70% per contributor majority agreement factor. Finally, the F1 method excludes contributors with lower then 100% objective term classification F1 scores. Each method results to a unique lexicon with varying emotional vectors for each term.

On applying the best majority filtering method to the 50% injection ratio, we noticed a remarkably high correlation. Due to the extensive filtering of the results, some methods are evaluated on a small subset of the total 456 terms. Figure 8 presents the IQR of per term correlation values between NRC and the results of the 50% objective ratio task. However, the high correlation of 'Majority' filtering is misleading. The number of terms - post filtering - was 46, which is almost a tenth of the original 456 terms. To better portray lexicon coverage, we assign an *Adjusted Score* to each term as follows:

$$AS = Spearman's\ Rho * \frac{Filtered\,terms}{Total\ terms} \qquad (7)$$

'Filtered terms' refers to the number of terms remaining after filtering, while *'Total terms*' is the number of terms used in each task -in our experiments: $Total terms = 456$. The correlation and the low coverage of Majority filtering is outlined in Table 3 column b in comparison to column a, $(a)\ 0.6498 * \frac{46}{456} = (b)\ 0.0656$.

Adjusted Score (AS) was consistently higher than 0.55 for every task and filtering method. The injection of objective terms improved the AS across all filtering methods, Table 4. In every task the F1 filtering presented the highest low whisker, $Q1 + 1.5 * IQR$. The upper quartile, Q3, was highest for best majority for every task. The majority that yielded the highest correlation with NRC was 70% for 50-33-25 injection ratios, Figure 9(a), 9(b) and 9(c), and 60% for the task with no injection, Figure 9(d). The best threshold was 30% for 50-33-0 injection ratios and 20% for the 25 injection ratio.

Correlation differences per task is relatively low. For 50% injection ratio F1 and Best Majority presents the highest median correlation. Best majority retains a high median correlation for 33% injection ratio, equal to Best Threshold. For the 25% and 0% ratios Best Majority presents the highest correlation. The variance is low for all methods, ranging from $9 * 10^{-5}$ to $4 * 10^{-4}$.

5.3. Emotional Diversity

The emotional diversity is defined as the multitude of annotated emotions per term. The set of Figures 10 presents the regression lines - with 95% confidence interval - of emotional diversity for each filtering method per injection ratio. The x-axis shows the number of different emotions in one term as per NRC, while the y-axis shows the number of different emotions in the same term post filtering.

The F1 filtering consistently provided a high number (> 2) of emotional diversity, Figures 10(a), 10(b) and 10(c). As the injection ratio is reduced the emotional diversity of F1 increased to up to 3 emotions per term.

Threshold filtering was strictly bound to the best performance threshold. When the 30% threshold was used, Figures 10(a),10(b) and 10(d), the number of emotions per term was higher than F1 filtering. However, when the best threshold was 20%, Figure 10(c), the emotions per term falls < 2. On the contrary, when the majority was stricter at 70%, the number of emotions per term was very low, Figures 10(a), 10(b) and 10(c). When the majority was set a lower 60% the emotional diversity increased.

Both threshold and majority filtering methods bound the distributions to their upper limits and directly affect the emotional distribution. Majority filtering was limiting diversity as it required single annotation agreement, while threshold filtering enforced diversity due to limiting peak class annotation.Our proposed F1 filtering is distribution agnostic, thus it doesn't directly alters the emotional diversity of each term.

5.4. Redundancy

Each filtering method had different redundancy and exclusion factors, Figures 11. F1 filtering maintained a redundancy higher than 6 for all injection ratios. As the injection ratio was decreasing, the redundancy level improved. A similar trend was noticed in the emotional diversity analysis, where lower injection ratios resulted in a higher number of emotion annotations. Conversely, Threshold filtering had an analogous to the injection ratio redundancy, probably because it was affected by the tight 20% threshold of the 25% injection ratio, Figure 11(a). Majority filtering had a redundancy lower than 5 throughout all the injection ratios. As the Majority filter lowers to 60%, for the 0% objective terms task, redundancy increases to ≈ 6.

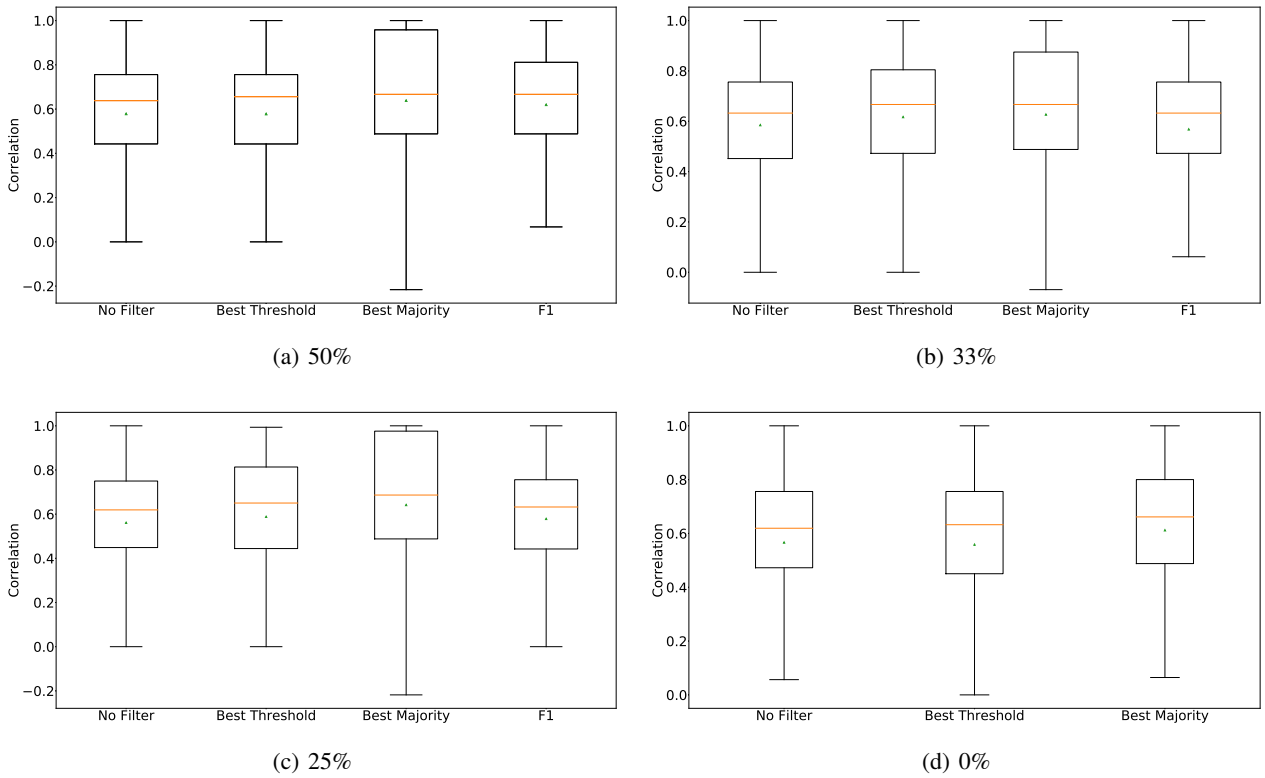
Nonetheless, the exclusion of annotations after filtering was significant, especially for Majority. High Majority requirements result in high exclusion. For all injection ratio the exclusion of annotations was higher than 60%, Figure 11(b). Strict threshold filtering increased exclusion, 25% injection ratio. F1 filtering exclusion was steadily lower than 40%.

6. Conclusions

Honest and non-spam contributions are of major importance for subjective tasks (Haralabopoulos et al., 2019;

	50%		33%		25%		0%	
Method	Mean	Median	Mean	Median	Mean	Median	Mean	Median
No filter	0.5781	**0.6381**	**0.5847**	0.6325	0.561	0.6193	0.5664	0.6193
Best Threshold	0.5777	0.656	**0.6167**	**0.6667**	0.588	0.6503	0.5585	0.6325
Best Majority	0.6379	0.6667	0.6268	0.6667	**0.6415**	**0.6865**	0.6117	0.6614
F1	**0.6191**	**0.6667**	0.5678	0.6325	0.5786	0.6325	*N/A*	*N/A*

Table 4: Mean and Median *Adjusted Score* correlation for different injection ratios



(a) 50%

(b) 33%

(c) 25%

(d) 0%

Figure 9: I.Q.R. of per term *Adjusted Score* for different objective term inclusion ratios

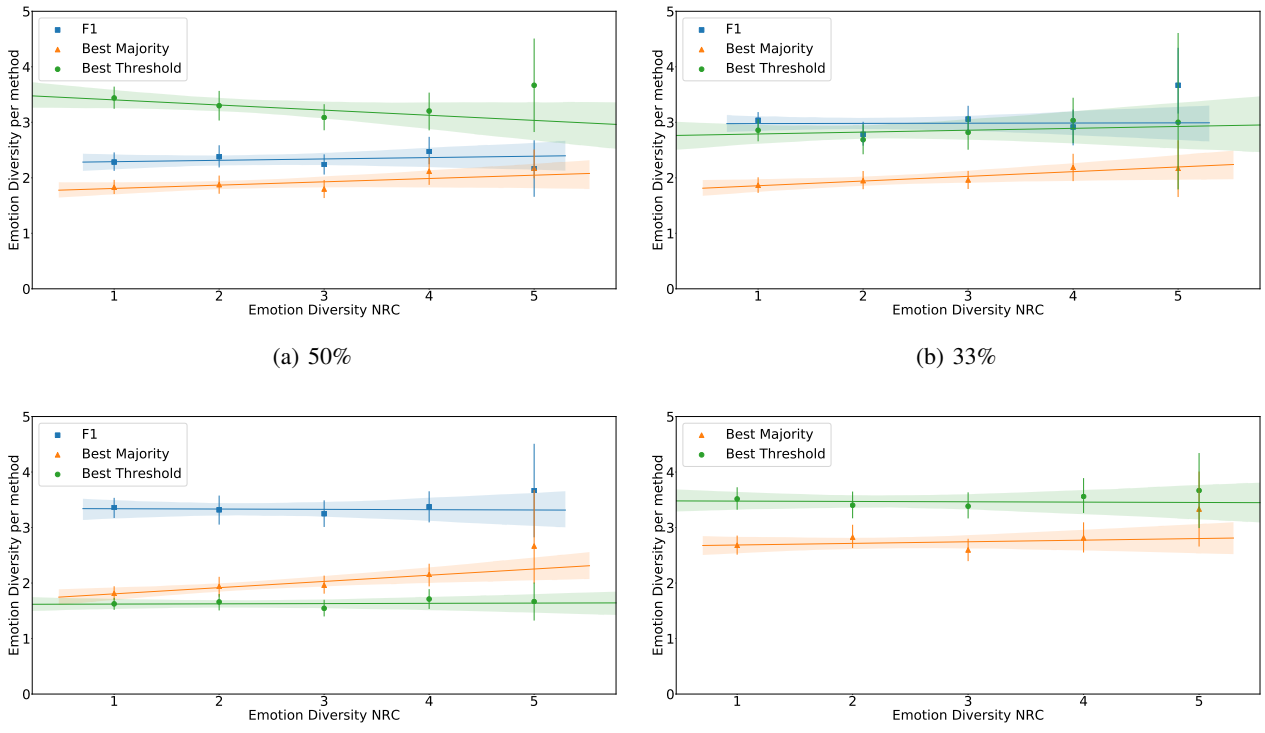
Jonell et al., 2018). We proposed an evaluation method based on objective terms and the evaluation of contributors based on a F1 contributor score, which is calculated only against the objective terms. The inclusion of objective terms and the filtering of dishonest or spamming annotations in a crowdsourcing task involves a direct resource cost. Requesters will need to allocated extra resources, to inject objective terms in addition to the desired subjective terms, to implement our proposed method. A varying level of injected terms is used to identify the trade-offs and costs of this filtering method.

We evaluated our proposed injection and the F1 filtering method with: correlation co-efficient analysis against an established lexicon, the analysis of the emotional diversity of the resulting terms, term redundancy and annotation exclusion ratio post filtering. Furthermore, we implemented two widely used filtering methods in crowdsourcing, Threshold and Majority, and calculated, based on their NRC correlation, the best performing filter bounds. The best Threshold and Majority filters, for each injection ratio, were also compared to our F1 filter.

Although we used NRC as the baseline for our evaluation, there were major emotional differences amongst the NRC lexicon and our annotation results, Figure 12. The NRC emotions of 'joy', 'fear', 'sadness' and 'anger' are outside the mean standard error range of our task results. Amongst those four emotions, 'joy' is marginalised in NRC when compared to our obtained emotional distributions. On the other hand, the intra-task correlation (0-25-33-50) is relatively high for all emotions. As we used a small subset (456 terms) from NRC, we cannot safely conclude whether the observed effects, of 'joy' suppression and emotional distribution difference, are lexicon-wide.

The inclusion of objective terms in the task improved the per term correlation irrespective of the filtering method. Our proposed F1 filtering method revealed a high correlation co-efficient with NRC, high emotional diversity, high redundancy and low exclusion ratio. F1 filtering improved all metrics when compared to the unfiltered results. Majority voting yielded the highest correlation results with low emotional diversity, low redundancy and high exclusion ratio. Finally, Threshold filtering had high correlation but was limited to the best performing threshold level on all three evaluations of diversity, redundancy and exclusion.

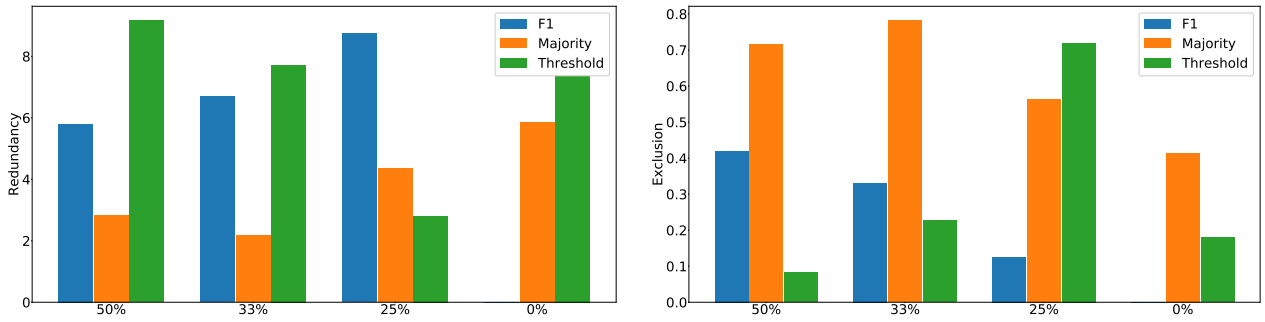
Most importantly, the contributor filtering of our approach doesn't directly affect the distribution of answers. A sub-



(a) 50% (b) 33%

(c) 25% (d) 0%

Figure 10: Emotional Diversity of filtered methods compared to NRC for different objective term inclusion ratios



(a) Redundancy (b) Exclusion

Figure 11: Mean redundancy per term (a) and annotation exclusion (b) for different injection ratios

jective task has no ground truth(Aroyo and Welty, 2015) and contributors should not be judged by their subjective contributions to the task. We instead provide an objective evaluation process suited to subjective tasks.

Going forward, we intend to evaluate the performance of our method in tasks with varying design and also expand to subjective sentence labelling. Our proposed objective evaluation method can: be used in any domain with domain specific objective terms for evaluation, assess high quality contributors and preserve subjectivity by excluding contributors with low evaluation scores but retaining all the quality annotations.

7. Funding

This research was funded by Engineering and Physical Sciences Research Council grant number EP/M02315X/1: ”From Human Data to Personal Experience”.

8. Bibliographical References

Aker, A., El-Haj, M., Albakour, M.-D., Kruschwitz, U., et al. (2012). Assessing crowdsourcing quality through objective tasks. In *LREC*, pages 1456–1461. Citeseer.

Aroyo, L. and Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Basile, V., Novielli, N., Croce, D., Barbieri, F., Nissim, M., and Patti, V. (2018). Sentiment polarity classification at evalita: Lessons learned and open challenges. *IEEE Transactions on Affective Computing*.

Figure 12: Mean Overall Annotations, for NRC and each injection ratio, with their respective standard error ranges

Brody, S., Keller, U., Degen, L., Cox, D. J., and Schächinger, H. (2004). Selective processing of food words during insulin-induced hypoglycemia in healthy humans. *Psychopharmacology*, 173(1-2):217–220.

Brosch, T., Pourtois, G., and Sander, D. (2010). The perception and categorisation of emotional stimuli: A review. *Cognition and emotion*, 24(3):377–400.

Budhi, G. S., Chiong, R., Hu, Z., Pranata, I., and Dhakal, S. (2018). Multi-pso based classifier selection and parameter optimisation for sentiment polarity prediction. In *2018 IEEE Conference on Big Data and Analytics (ICBDA)*, pages 68–73. IEEE.

Calefato, F., Lanubile, F., and Novielli, N. (2017). Emotxt: a toolkit for emotion recognition from text. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 79–80. IEEE.

Canetti, L., Bachar, E., and Berry, E. M. (2002). Food and emotion. *Behavioural processes*, 60(2):157–164.

Caselli, T., Sprugnoli, R., and Inel, O. (2016). Temporal information annotation: crowd vs. experts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3502–3509.

Chamlertwat, W., Bhattarakosol, P., Rungkasiri, T., and Haruechaiyasak, C. (2012). Discovering consumer insight from twitter via sentiment analysis. *J. UCS*, 18(8):973–992.

Chaplin, T. M. (2015). Gender and emotion expression: A developmental contextual perspective. *Emotion Review*, 7(1):14–21.

Chaturvedi, I., Cambria, E., Welsch, R. E., and Herrera, F. (2018). Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*, 44:65–77.

Diakopoulos, N. A. and Shamma, D. A. (2010). Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1195–1198. ACM.

Eickhoff, C. (2018). Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 162–170. ACM.

Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124.

Ekman, P. and Keltner, D. (1997). Universal facial expressions of emotion. *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture*, pages 27–46.

Elfenbein, H. A. and Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203.

Elfenbein, H. A. and Luckman, E. A. (2016). 16 interpersonal accuracy in relation to culture and ethnicity. *The Social Psychology of Perceiving Others Accurately*, page 328.

Elfenbein, H. A. (2017). Emotional dialects in the language of emotion. *The science of facial expression*, pages 479–496.

Ellis, C. and Flaherty, M. G. (1992). An agenda for the interpretation of lived experience. *Investigating subjectivity: Research on lived experience*, pages 1–13.

Fernández-Gavilanes, M., Juncal-Martínez, J., García-Méndez, S., Costa-Montenegro, E., and González-Castaño, F. J. (2018). Creating emoji lexica from unsupervised sentiment analysis of their descriptions. *Expert Systems with Applications*, 103:74–91.

Finnerty, A., Kucherbaev, P., Tranquillini, S., and Con-

vertino, G. (2013). Keep it simple: Reward and task design in crowdsourcing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*, page 14. ACM.

Ghosal, D., Akhtar, M. S., Ekbal, A., and Bhattacharyya, P. (2018). Deep ensemble model with the fusion of character, word and lexicon level information for emotion and sentiment prediction. In *International Conference on Neural Information Processing*, pages 162–174. Springer.

Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., and Reyes, A. (2015). Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–478.

Gohier, B., Senior, C., Brittain, P., Lounes, N., El-Hage, W., Law, V., Phillips, M. L., and Surguladze, S. (2013). Gender differences in the sensitivity to negative stimuli: Cross-modal affective priming study. *European Psychiatry*, 28(2):74–80.

Haralabopoulos, G. and Simperl, E. (2017). Crowdsourcing for beyond polarity sentiment analysis a pure emotion lexicon. *arXiv preprint arXiv:1710.04203*.

Haralabopoulos, G., Wagner, C., McAuley, D., and Simperl, E. (2018). A multivalued emotion lexicon created and evaluated by the crowd. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 355–362. IEEE.

Haralabopoulos, G., Wagner, C., McAuley, D., and Anagnostopoulos, I. (2019). Paid crowdsourcing, low income contributors, and subjectivity. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 225–231. Springer.

Hazarika, D., Poria, S., Vij, P., Krishnamurthy, G., Cambria, E., and Zimmermann, R. (2018). Modeling interaspect dependencies for aspect-based sentiment analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 266–270.

Hetmank, L. (2013). Components and functions of crowdsourcing systems-a systematic literature review. *Wirtschaftsinformatik*, 4:2013.

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, 14(6):1–4.

Hutto, C. J. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Jonell, P., Oertel, C., Kontogiorgos, D., Beskow, J., and Gustafson, J. (2018). Crowdsourced multimodal corpora collection tool. In *The Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 728–734.

Kang, D. and Park, Y. (2014). based measurement of customer satisfaction in mobile service: Sentiment analysis and vikor approach. *Expert Systems with Applications*, 41(4):1041–1050.

Keith, K. D. (2019). *Cross-cultural psychology: Contemporary themes and perspectives*. John Wiley & Sons.

Kiritchenko, S. and Mohammad, S. M. (2017). Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling. *arXiv preprint arXiv:1712.01741*.

Kittur, A., Chi, E. H., and Suh, B. (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM.

Knapp, M. L., Hall, J. A., and Horgan, T. G. (2013). *Nonverbal communication in human interaction*. Cengage Learning.

Koltsova, O. Y., Alexeeva, S., and Kolcov, S. (2016). An opinion word lexicon and a training dataset for russian sentiment analysis of social media. *Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE 2016 (Moscow)*, pages 277–287.

Lau, R. Y., Li, C., and Liao, S. S. (2014). Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis. *Decision Support Systems*, 65:80–94.

Lauka, A., McCoy, J., and Firat, R. B. (2018). Mass partisan polarization: Measuring a relational concept. *American behavioral scientist*, 62(1):107–126.

Li, G., Wang, J., Zheng, Y., Fan, J., and Franklin, M. J. (2018). Crowdsourcing background. In *Crowdsourced Data Management*, pages 11–20. Springer.

Luhrmann, T. M. (2006). Subjectivity. *Anthropological Theory*, 6(3):345–361.

Masuda, T., Gonzalez, R., Kwan, L., and Nisbett, R. E. (2008). Culture and aesthetic preference: Comparing the attention to context of east asians and americans. *Personality and Social Psychology Bulletin*, 34(9):1260–1275.

Matsumoto, D., Hwang, H. C., and Frank, M. G. (2013). Emotional language and political aggression. *Journal of Language and Social Psychology*, 32(4):452–468.

Maynard, D. and Bontcheva, K. (2016). Challenges of evaluating sentiment analysis tools on social media. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1142–1148. LREC.

Miao, H., Liu, R., Gao, S., Zhou, X., and He, X. (2018). End-to-end deep memory network for visual-textual sentiment analysis. In *2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, pages 399–403. IEEE.

Mihalcea, R., Banea, C., and Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 976–983.

Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

O'Leary, D. E. (2016). On the relationship between number of votes and sentiment in crowdsourcing ideas and comments for innovation: A case study of canada's digital compass. *Decision Support Systems*, 88:28–37.

Öztürk, N. and Ayvaz, S. (2018). Sentiment analysis on

twitter: A text mining approach to the syrian refugee crisis. *Telematics and Informatics*, 35(1):136–147.

Palan, S. and Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Peer, E., Brandimarte, L., Samat, S., and Acquisti, A. (2017). Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163.

Pessoa, L. and Adolphs, R. (2010). Emotion processing and the amygdala: from a'low road'to'many roads' of evaluating biological significance. *Nature reviews neuroscience*, 11(11):773.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Prabowo, R. and Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143–157.

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., and Mozetič, I. (2015). The effects of twitter sentiment on stock price returns. *PloS one*, 10(9):e0138441.

Reidy, D. E., Brookmeyer, K. A., Gentile, B., Berke, D. S., and Zeichner, A. (2016). Gender role discrepancy stress, high-risk sexual behavior, and sexually transmitted disease. *Archives of sexual behavior*, 45(2):459–465.

Rouder, J., Morey, R., and Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra: Psychology*, 2(1).

Schouten, K., Van Der Weijde, O., Frasincar, F., and Dekker, R. (2018). Supervised and unsupervised aspect category detection for sentiment analysis with co-occurrence data. *IEEE transactions on cybernetics*, 48(4):1263–1275.

Schumaker, R. P., Jarmoszko, A. T., and Labedz Jr, C. S. (2016). Predicting wins and spread in the premier league using a sentiment analysis of twitter. *Decision Support Systems*, 88:76–84.

Semnani-Azad, Z. and Adair, W. L. (2013). Watch your tone… relational paralinguistic messages in negotiation: The case of east and west. *International Studies of Management & Organization*, 43(4):64–89.

Sharma, S. and Chakraverty, S. (2018). An approach to track context switches in sentiment analysis. In *Progress in Advanced Computing and Intelligent Engineering*, pages 273–282. Springer.

Silvers, J. A., Insel, C., Powers, A., Franz, P., Helion, C., Martin, R. E., Weber, J., Mischel, W., Casey, B., and Ochsner, K. N. (2016). vlpfc–vmpfc–amygdala interactions underlie age-related differences in cognitive regulation of emotion. *Cerebral Cortex*, 27(7):3502–3514.

Smith, A. K. and Elias, L. J. (2018). Native reading direction modulates eye movements during aesthetic preference and brightness judgments. *Psychology of Aesthetics, Creativity, and the Arts*.

Smith, J. A. (2015). *Qualitative psychology: A practical guide to research methods*. Sage.

Starr, M. S. and Rayner, K. (2001). Eye movements during reading: Some current controversies. *Trends in cognitive sciences*, 5(4):156–163.

Tang, D., Qin, B., and Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432.

Teo, T. (2018). Homo neoliberalus: From personality to forms of subjectivity. *Theory & Psychology*, 28(5):581–599.

Valdivia, A., Luzón, M. V., Cambria, E., and Herrera, F. (2018). Consensus vote models for detecting and filtering neutrality in sentiment analysis. *Information Fusion*, 44:126–135.

White, S. J., Warrington, K. L., McGowan, V. A., and Paterson, K. B. (2015). Eye movements during reading and topic scanning: Effects of word frequency. *Journal of Experimental Psychology: Human Perception and Performance*, 41(1):233.

Yao, Y.-W., Liu, L., Ma, S.-S., Shi, X.-H., Zhou, N., Zhang, J.-T., and Potenza, M. N. (2017). Functional and structural neural alterations in internet gaming disorder: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, 83:313–324.

Yoshino, K., Ishikawa, Y., Mizukami, M., Suzuki, Y., Sakti, S., and Nakamura, S. (2018). Dialogue scenario collection of persuasive dialogue with emotional expressions via crowdsourcing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Yue, L., Chen, W., Li, X., Zuo, W., and Yin, M. (2018). A survey of sentiment analysis in social media. *Knowledge and Information Systems*, pages 1–47.

Zamil, A. A. A., Hasan, S., Baki, S. M. J., Adam, J. M., and Zaman, I. (2019). Emotion detection from speech signals using voting mechanism on classified frames. In *2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, pages 281–285. IEEE.

Zhao, W., Guan, Z., Chen, L., He, X., Cai, D., Wang, B., and Wang, Q. (2018). Weakly-supervised deep embedding for product review sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 30(1):185–197.