A multivalued emotion lexicon created and evaluated by the crowd

Giannis Haralabopoulos, Christian Wagner, Derek McAuley
University of Nottingham
name.surname@nottingham.ac.uk
Elena Simperl
University of Southampton
E.Simperl@soton.ac.uk

***Abstract*—Sentiment analysis aims to uncover emotions conveyed through information. In its simplest form, it is performed on a polarity basis, where the goal is to classify information with positive or negative emotion. Recent research has explored more nuanced ways to perform emotion analysis. Unsupervised emotion analysis methods require a critical resource: a lexicon that is appropriate for the task at hand, in terms of the emotional range and diversity captured. Emotion analysis lexicons are created manually by domain experts and usually assign one single emotion to each word. We propose an automated workflow for creating and evaluating a multi- valued emotion lexicon created and evaluated through crowdsourcing. We compare the obtained lexicon with established lexicons and appoint expert English Linguists to assess crowd peer-evaluations. The proposed workflow provides a quality lexicon and can be used in a range of text property association tasks.**

Keywords: Crowdsourcing, Beyond Polarity, Pure Emotion, Sentiment Analysis, Lexicon Acquisition, Reddit, Twitter

I. Introduction

Sentiment analysis aims to uncover the emotion conveyed through information, based on a set of methods (NLP, rule/frequency based, similarity measurement). In online social networks, sentiment analysis is mainly performed for political and marketing purposes, product acceptance and feedback systems. This involves the analysis of various social media information types, such as text [34], emoticons and hashtags, or multimedia [51]. However, to perform sentiment analysis, information has to be labelled with a sentiment. This relationship is defined with a lexicon.

Lexicon acquisition is a requirement for unsupervised sentiment classification. During the acquisition process, individual or grouped information elements are labelled based on a selections of classes. Sentiment classification is the task that uses the acquired lexicon and a classification method to classify a sentence, phrase, or social media submission as a whole, based on the aggregation of its terms' labels. Thus, lexicon quality directly affects sentiment classification accuracy.

Both tasks can either be performed automatically [22] or manually [33], where the labelling is done by linguists or researchers themselves [1]. Apart from experts, manual labelling can also be performed with the help of a wide network of people, known as crowdsourcing [31]. Crowdsourcing is widely used for polarity lexicons, but rarely for beyond-polarity and - to the best of the authors' knowledge, so far has never been for the discovery of other linguistic elements such as intensifiers, negators, or stop words [55, 18].

Sentiment analysis is commonly performed on a polarity basis, i.e. the distinction between positive and negative emotion. These poles correspond to agreement and disagreement, or acceptance and disapproval, for candidates and products respectively [65]. Beyond-polarity or emotion sentiment analysis aims to uncover an exact emotion, as defined by emotional theories and physiologists [48, 21]. Emotion analysis studies most frequently acquire lexicons based on the evaluation of experts and use a single emotion per term [11]. Natural Language Processing (NLP) applications that rely on experts are less comprehensive, and not as scalable, compared to crowdsourced NLP applications [23].

A. Motivation

Existing emotion lexicons have strengths and weaknesses according to their design and result processing. Single valued lexicons are usually driven by a gold standard, which essentially removes collected annotations. Every annotation in disagreement with the gold standard is discarded, resulting in data loss. However, authors of [6] note that there is no truth in human intelligence tasks, more so on subjective ones like emotion annotation. Therefore a gold standard might not reflect the truth, but only portrait a personalised truth as defined by the expert(s) employed. Existing multivalued emotion lexicons, such as [41], assign binary values to emotions via a consolidation method that makes scalability difficult if not impossible. Our proposed lexicon includes the exact annotations simplifying scalability.

We propose a crowd-centric multivalued emotion lexicon acquisition process, based on Plutchik's eight basic emotions [48], that is scalable and cost effective. The crowd performs the annotation, identifies linguistic elements, as opposed to pooling them from existing lists [36], and evaluates the annotations provided. Crowd evaluations are compared to domain expert evaluations in order to assess crowd capabilities of evaluating term-emotion associations. The created lexicon is then compared to the established NRC lexicon, to assess its overall quality. The workflow presented can be applied in multiple domains as well, while text property association tasks can benefit from its multivalued approach. Examples of text property association tasks are: medical records and medical conditions association, social media submissions and probabilistic recommendations, work environment correspondence and feedback systems. Our lexicon is provided as is for emotion classification tasks, and the proposed workflow as a base for building application specific lexicons without employing experts.

II. Background

According to [15], an emotion is defined with reference to a list. Ekam et al. [21] proposed the six basic emotions joy, anger, fear, sadness, disgust, and surprise. Years later, Plutchik [48, 49] proposed the addition of trust and anticipation as basic emotions, and presented a circumplex model of emotions, which defines emotional contradictions and some of the possible combinations.

Sentiment analysis aims to classify information based on the emotion conveyed. Depending on the number of classes/emotions required, we can separate the analysis into: polarity and beyond-polarity. Polarity sentiment analysis studies define two opposite emotional states, positive and negative, or good and bad with the addition of a neutral state [45, 63, 3]. Furthermore, some researchers have classified information on levels for each pole(e.g. very positive, positive, neutral, negative, very negative etc.), also known as fine grained sentiment analysis [62, 25, 57].

Emotion analysis, also known as beyond-polarity or pure emotion,is a refined sentiment analysis, that incorporates a wider range of possible emotion labels. Examples of emotional labels might be –but are not limited to– : sadness, boredom, joy, sadness, surprise, anger, fear, disgust etc. [41, 20, 52, 64].

As discussed in Section 1, one of the core tasks of text based sentiment analysis is lexicon acquisition. A lexicon can be acquired through manual or automatic annotation. However, natural language has a very subjective nature [4] which significantly inhibits automated sentiment lexicon acquisition methods from achieving relevance equal to manual methods [38]. Thus a lot of researchers choose to manually annotate their term corpora [50, 19], or use established lexicon such as WordNet [35, 10, 54, 58] and SentiWordNet [9, 53, 7], or other lexicons [30, 25, 47]. Other studies combine manual labelling or machine learning with lexicons [46].

Manual lexicon acquisition is constrained by the number of people contributing to the task, and the number of annotations from each participant. These constraints can be eliminated by increasing the number of people involved, for instance, by using crowdsourcing [14]. Amazon's Mechanical Turk (MTurk)[1] is a crowdsourcing platform frequently used for polarity sentiment lexicon acquisition via crowdsourcing [31, 37, 32, 44]. MTurk has also been used, for the annotation of one thousand tweets in [20], more than ten thousand terms [41], and the annotation of ninety five emoticons out of one thousand total emoticons found in [64]. Authors of [52] had one thousand four hundred terms labelled with a supervised machine learning and crowd validators.

The second core part in sentiment analysis, is sentiment classification –a classification that occurs at phrase/sentence/submission level, and is usually based on the aggregation of the term's labelled emotions. As with lexicon acquisition, the classification task can be automated [24, 25, 28, 60, 17, 59] or performed manually [29, 43, 26, 13].

Regardless of manual or automated sentiment classification, on textual information scenarios, term/phrase sentiment is the input of the classification process. In some cases the appointed class might be different from the individual term/phrase emotion, leading to relabelling of the terms [56]. Manually labelled classification can achieve high accuracy, but it requires additional resources, and is not easily scalable. On the other hand, automated processes are scalable but have lower accuracy [13, 27].

III. Proposed Methodology

The proposed methodology is comprised of the data collection process, the suggested workflow, followed by the analysis of the results and their evaluation.

A. Data Collection

Data collection is an integral part of lexicon creation. Modern text analysis is moving towards social networks transcripts, thus anonymised social media submissions are a good resource for term acquisition. The diversity of participants in an online network provides a mix of formal and informal text submissions. Furthermore, controversial topics highlight the need for a multivalued approach of lexicons, as the emotional responses are more diverse than non-controversial terms.

The labelling process is performed by anonymous crowd contributors. It is suggested to employ voluntary crowd contributors as the quality of their contributions is higher than those that participate in monetary incentivised tasks [39]. However, the time to complete the task when utilising voluntary contributors is significantly longer [12]. When dealing with unigrams there is no need for anonymisation, but when dealing with n-grams ($n > 1$) content has to be anonymised to preserve the identity of social media users.

B. Workflow

Our proposed workflow is comprised of 3 core processes: Data preprocessing, Labelling and Evaluation, Figure 1. Data preprocessing is automated, while Labelling and Evaluation are crowdsourced.

The first core process requires a text collection and includes the discovery of its underlying properties. Depending on the nature of the research different textual properties and forms needed, e.g. sentiment in a sentiment analysis study, lemmas or syntactic function in a linguistic study. A plethora of libraries exist, in various programming environments, that can automatically process text to the desired form, e.g. in our study we are interested in stems and unigrams.

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

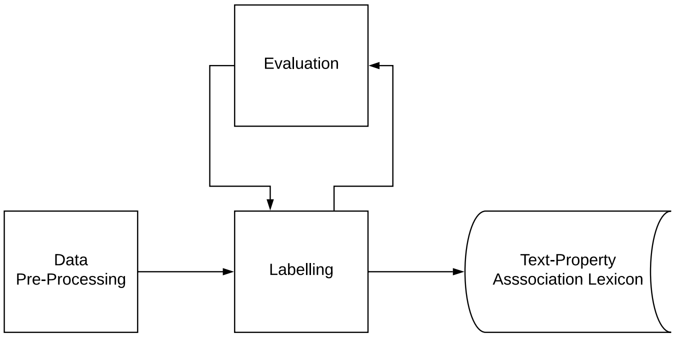

Fig. 1. Lexicon creation workflow

The second core process is manual labelling. Human annotators identify the text properties and provide the corresponding labels. As expert annotators are hard to find and often cost a lot to employ, we employed non expert annotators for the labelling process. Human annotation is a subjective task and a range of labels is produced, instead of a unique label desired in objective tasks, therefore we propose the storage of all the annotation information provided by contributors. Human annotators are also tasked with identifying linguistic elements, which in our study are stop words, intensifiers and negators.

The evaluation of crowdsourcing tasks is usually performed by experts. In our workflow crowd contributors evaluate their peers. The quality of the labels can be used as a feedback for crowdsourcing, e.g. low quality can act as a marker for higher redundancy. The proposed workflow doesn't require any experts to filter the annotations, set the ground truth or evaluate the results.

The lexicon is stored with its individual annotations, which provides a multi-property text association and enhances scalability. Researchers are able to apply consolidation or majority selection methods if a single property association as suited to their needs.

C. Analysis

The obtained results will be analysed under a diversity scope. On the specific topic of emotions, emotional combinations can be interpreted to different emotions, thus emotional diversity conveys more information [48]. On the contrary, annotational agreement is a forced prerequisite which results in restricted emotional interpretations.

D. Evaluation

The resulting lexicon are peer evaluated. Crowd contributors evaluate the obtained annotations per term with a scaling responses method. To our best knowledge this is the first time crowd contributors evaluate their peers. Thus, domain experts are employed to assess the crowd capabilities in evaluating contributions. In addition, since a similar lexicon exists, we compare term inclusions and annotations of both lexicons.

IV. Demonstration

A. Data Collection

During January 2017, we performed a keyword based crawl for articles and comments in the Europe subreddit[2] and Twitter[3] tweets that contained the word ”Brexit”. The use of a political and controversial term in the query is deliberate, to capture the emotional diversity of political statements. We crawled one hundred articles from Reddit, with more than forty thousand comments and more than three thousand tweets. In total, the number of unique terms in our corpus is 30227, more than 19 thousands of them were validated with a Great British English dictionary [2, 16]. The validated terms follow Zipf's Law [66] with scaling-law coefficient $a = 1$.

The crowdsourcing task, hosted in Figure-Eight[4], required contributors to label terms in three different main classes, *emotion*, *intensifier* and *none*. Emotion labelling included the 8 basic emotions as defined by Plutchik. Intensifier class included intensifiers and negators, and *none* referred to stop words or words with no particular emotion.

More than one hundred eighty contributors performed eighty thousand annotations. Most of the contributors annotated the maximum allowed number of term groups, 1% of the total annotations needed. The simplicity of the task resulted in high overall contributor engagement, with 429 mean and 580 median annotations per contributor. The task was completed within 7 hours.

B. Workflow

Our goal is to create an end to end automated workflow for the creation, evaluation and enrichment of text-property association lexicons.

Data preprocessing is comprised of 3 unsupervised steps: tokenisation, stemming and spell check. Textual content is term tokenised, terms are then checked for spelling and stemmed based on their root. The resulting stems along with their stem groups are stored as a single entry. Term grouping might alter the emotional properties of contained terms, but reduces costs, time required and provides a range of benefits to machine learning applications [8]. The tools for this core part were developed in Python using (amongst others) the enchant library[5] and Natural Language Toolkit[6].

Crowdsourcing acts as an always available human computation unit that provides text property association information, emotions in our study. The task requires contributor to choose a main class, *emotion*, *intensifier* and *none*, and a subclass. The subclasses are the eight emotions, the type of intensifier and none. Each of the eleven options for subclasses, will be referred to as ”subclass”. To assist contributors with term definitions, every term group had a link to an English dictionary.

Crowd annotations define (a) main subclass(es), which refers to the subclass(es) that received the majority of annotations, subclass annotations refer to other subclass(es) annotated

[2] https://www.reddit.com/r/europe/
[3] https://twitter.com/search-home
[4] https://www.figure-eight.com/
[5] https://pypi.python.org/pypi/pyenchant/
[6] https://www.nltk.org/

from the contributors. Two or more main subclasses occur on annotation agreement, i.e. when the number of annotations for two or more subclasses are the same. Spam filtering and annotation quality measures are utilised for contribution quality purposes.

The performance and the quality of the human computation unit is monitored via peer evaluation. Crowd contributors, different than the ones participated in the annotation task, evaluate the annotations based on a summary of the annotations received per the term group. The evaluation is performed on a validity scale from 1 to 5. The subjective nature of the task fits better under a validity scope, rather than the scope of correctness. Peer evaluation for crowdsourcing is largely unexplored. To assess the efficiency and the applicability of a peer design in crowdsourcing, we compare the evaluations of the crowd to the evaluation of two –unaffiliated to the authors– Post-Doctoral English Linguists.

C. Analysis

The text-emotion lexicon (will be referred as simply ”lexicon”) is created after spam and quality filtering of the received sixty thousand annotations. Terms in our lexicon are grouped based on their stem. This resulted in a 40% reduction of the initial single term corpus. Stemming significantly reduces cost and time-required for the task.This initial version of the lexicon contained more than twenty thousand annotations for 9737 term groups. Each term group received a mean 2.3 annotations from a total of 95 different annotators. Although the number of mean annotations in the final lexicon is less than half the mean annotations in the unfiltered corpus, the remaining annotations should be considered of honest (if not of higher quality) based on the filtering processes employed.

TABLE I
SAMPLE OF NON-EMOTIONAL ANNOTATED TERM GROUPS

Intensifiers	**Negators**	**None**
harder	dispensation dispense	is
largely large	minimize minimal	because
mostly	eliminates eliminated	to

Most of term groups in the lexicon have diverse subclass annotations. The dominant emotion in our lexicon is *joy*, while the least annotated emotion is *disgust*. Additionally, 148 terms were annotated as intensifiers, 43 terms as negators, and 6801 terms as *none*. A sample of term groups for each of the non emotional subclasses can be seen in Table I. The full lexicon can be found at Github[7] with detailed instructions[8].

Intensifiers and negators serve as modifiers to the emotional context of a term. Contributors identified intensifiers and negators that can modify emotion evoking words in the absence of context. Based on the received annotations there is room for improvement on the description of the structural role of intensifiers and the provided examples, as a number of non intensifying words were wrongfully annotated. The intensifier class contradicts the overall subjective nature of the emotional annotation, but presents the capabilities of the crowd in identifying the whole range of elements in a text-property association and mixed purpose tasks.

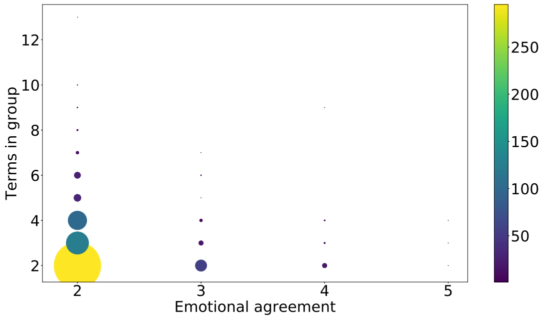

Fig. 2. Size and colour indicate number of term groups with emotional agreement

Annotation agreement refers to equal number of annotations in multiple subclasses or emotions. The vast majority of term groups in our lexicon doesn’t display any form of contradicting annotation agreement. Contradicting emotions and emotional combinations as described in [48] appear in the lexicon, but only 21% and 20% of the term groups had a subclass and an emotional agreement respectively. Contradicting or multi-emotion agreement was observed in 8.6% of the total term groups.

The number of subclasses in agreement and the number of terms in a term group are negatively correlated. Term groups with two terms appeared to have the highest subclass agreement with exactly two subclasses. The most common occurring agreements were subclass *none* paired with an emotion, and *joy* paired with an emotion. The number of multi-class agreement occurrences was disproportional to the number of terms in a term group. This is a strong indication that stemming didn’t confuse contributors.Similarly, the number of emotional agreement is disproportionate to the number of terms in the term group Figure 2. Furthermore, emotional agreement appeared in 10% of the term groups, while subclass agreement was found in 20% of the term groups.

In the agreement annotations, *joy* is the most common emotion. As previously mentioned, according to Plutchik each emotion has a contradicting one, and pairs of emotions indicate a more ”complex” emotion. There are 697 emotional agreeing term groups, of 1434 terms, with exactly two emotions. These emotional dyads[48] can be combined as seen in Table II. Simple basic emotion annotation tasks can indirectly provide complex emotional annotations.

TABLE II
SAMPLE OF COMBINATION DYADS

Dyad	**Emotion**	**Term groups**	**Terms**
trust joy	love	94	231
joy anticipation	optimism	58	142
surprise joy	delight	43	88
fear joy	guilt	39	89

Dyadic emotional agreements could be interpreted as the

[7] https://raw.githubusercontent.com/GiannisHaralabopoulos/Lexicon/master/lexicon.csv

[8] https://github.com/GiannisHaralabopoulos/Lexicon

resulting complex emotion, or further annotated to obtain a single dominant emotion. There was a number of term groups with opposite emotion dyads, presented in Table III,but as the number of annotations increases, emotional agreement occurrences -combination or opposition- decreases.

TABLE III
OPPOSITION DYADS

Dyad	Term groups	Terms
sadness joy	55	90
anger fear	20	34
surprise anticipation	16	30
disgust trust	12	18

In total, the lexicon features 17740 annotated terms with 3 classes and 11 subclasses.The dominant class for 7030 terms was *emotion*, 191 *intensifying*, 6801 *none*, and 3718 in some form of subclass agreement. Lexicon terms are mainly *joy* annotated, and emotional agreement is prevalent in 10% of the terms. Only 21% of total terms have a subclass agreement.

D. Evaluation

The lexicon is evaluated from the crowd with a Likert-type scale of validity. Crowd evaluations are compared to those of two Post-Doctoral English Linguists to assess the crowd's capabilities in peer evaluation. Moreover, the lexicon is compared with the an existing multivalued emotional lexicon.

Experts

We perform a direct comparison of expert and crowd evaluation. Crowd evaluation is a main part of our workflow, but peer evaluation in crowdsourcing is unexplored. Therefore we need to assess the evaluation capabilities of the crowd against the established evaluation by experts. We decide not to evaluate the lexicon based on a single emotion chosen but instead use a Likert-type scale of validity.

We sampled 1000 term groups based on the number of total annotations (200 term groups for each number of annotations from 2 to 6). The experts are two Post Doctoral English linguists unaffiliated to the authors, while the crowd is made up of contributors that choose to participate in the task. The cost of hiring these two experts is equal to the cost of employing nineteen contributors in Figure-Eight platform.

Evaluators were given a summary of the annotations received for one term group in the form of:*The term group "inequality inequity" received annotations as 50.0% sadness, 33.33% disgust, 16.67% anger.* Then, they were asked to evaluate, on a scale from 1 to 5, how valid these annotations were considered. The validity measurement will refer to the mean score from all the evaluations received, for both experts and crowd.

The summary of the evaluation can be seen in Figure 3. The first graph presents the validity over the number of annotations in the main class of the term group, where high annotational agreement corresponds to high evaluation scores. Both experts and the crowd follow that positive trend. Crowd contributors are more strict in their evaluations, but after four annotations we observe a significant validity increase on both crowd and experts.

Likewise, the annotation percentage for the majority class has a positive influence to the evaluation score, with the exception of 100% agreement, second graph Figure 3. The weighting factor for term groups with 100% annotation agreement is the reduced number of total annotations. On term groups with low number of total annotations, agreement is more prevalent.

In emotion annotations, as seen on the third graph of Figure 3, crowd and experts follow a similar evaluation pattern. *Anticipation* and *joy* had the exact same evaluation, while every other emotion and stop words were evaluated lower from the crowd. The only subclasses evaluated higher from the crowd were intensifiers and negators, with a significant difference in the evaluations for the latter. Section 6.3 provides a more detailed evaluation for term groups that received at least one annotation as intensifiers or negators.

The final graph in Figure 3 presents a clear negative correlation of subclass agreement and evaluation scores. The highest number of subclasses that do not affect evaluation scores is three, above that there is a steady decline of the evaluation scores, from both the crowd and the experts.

This direct comparison provides some insights on the crowd and expert evaluation capabilities and performance. On all occasions, except negation annotated terms, experts evaluated terms with higher validity than the crowd. Expert and crowd evaluations follow the same positive or negative correlations. We believe that the results are a fair indicator of the crowd's capabilities in peer evaluation.

The results highlight the importance of redundancy in crowdsourcing. Annotational agreement and majority voting are important, but validity remains between 3 and 4, from 25% to 80% majority for expert and crowd evaluations, and from 25% to 100% for crowd evaluations. Subclass agreement has a negative effect on three or more subclasses. Most importantly and compared to experts, the crowd is a stricter evaluator, that leads to higher quality annotations [5], with significantly lower costs, and higher scalability. Crowd contributors can be found in high numbers, multiple platforms, and with lower costs compared to expert linguists.

1) Intensifiers and negators: The task of evaluating intensifiers and negators was similar to the emotional annotation evaluation. Crowd and experts were evaluating each term group on the inclusion of at least one valid intensifier or negator. We used 541 term groups from the lexicon that had at least one annotation in any of the intensifying subclasses. Although, the particular selection of term groups is statistically significant, we expect relatively low evaluation scores as there are terms groups with minor annotations as intensifiers or negators. The term groups with majority annotations in intensifying class were less than 20.

In Figure 4, we define varying levels of agreement on the validity of the intensifying class, based on the agreement of evaluators. For the experts group, *low agreement* refers to term groups that received at least one out of two evaluations as valid, while *high agreement* requires the evaluation agreement of both experts. Similarly for the crowd, *low agreement* refers to a minimum of two valid evaluations, *mid agreement*

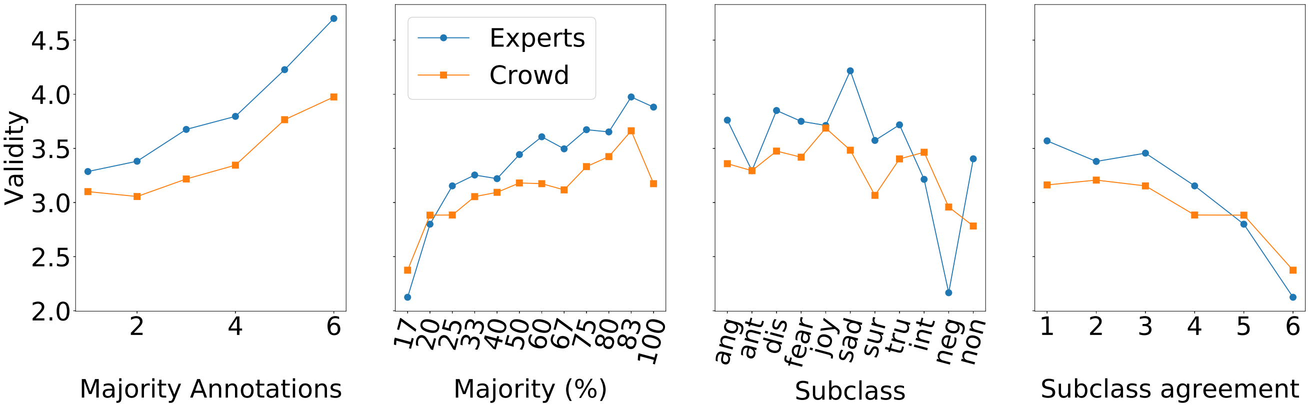

Fig. 3. Validity over: Majority subclass annotations, Majority subclass annotations percentage, Subclasses, Subclass Agreement

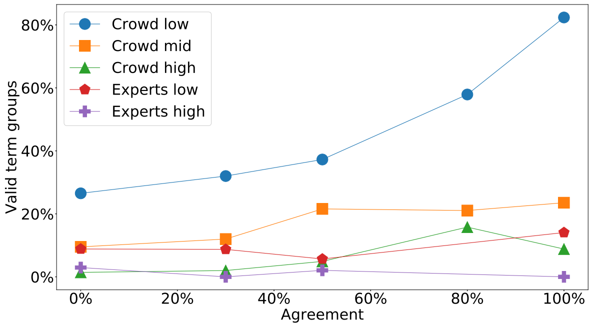

Fig. 4. Intensifying class evaluations

corresponds to three, and *high agreement* requires an absolute agreement of all four evaluators.

Experts are far more strict than the crowd in the evaluation of intensifiers and negators. When the validity agreement is low on both evaluation groups, the average valid term group difference is more than 40%, but the high validity agreement the difference is just 5.33%. When high agreement evaluation is applied, the crowd and expert evaluations are almost identical. The number of evaluations provides a degree of freedom in the evaluation strictness.

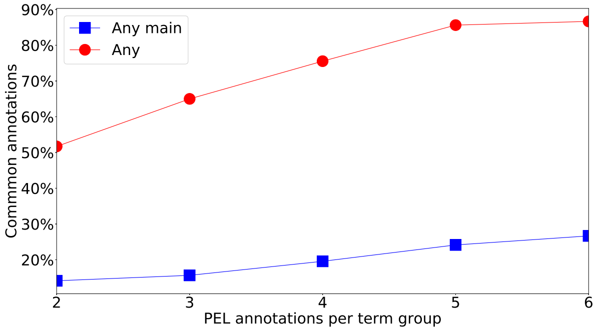

Fig. 5. Common classes for NRC and PEL lexicons

Comparison with other lexicons

We compare the lexicon with the widely known NRC word emotion association lexicon [42]. To maintain uniformity with our lexicon creation process, terms from NRC lexicons are checked with enchant python library[9] and stemmed with Porter Stemming Algorithm [61] from NLTK library[10]. The stemming process creates term groups, from NRC terms of the same root, in line with the term groups found in our lexicon. NRC terms have binary emotional annotation which are added up, as part of the same group, to create a comparable lexicon structure.

Out of the total 3716 term groups in NRC, the number of common emotional term groups with our lexicon is 2412. The number would be higher, if NRC word emotion association lexicon included non-emotional terms. We can consider the highest emotional annotation as main class and the rest as subclasses on both lexicons. This formulation gives us the ability to perform a direct comparison of the common term group annotations, Figure 5. *Any main* requires at least one similar main class across lexicons, for term group that have multiple emotional main classes. While *Any* refers to at least one common emotional annotation among NRC and lexicon term groups.

Overall, the number of annotations per term group is proportional to the common lexicon classes. The distribution of emotions spreads over multiple emotions as annotations increase, thus the huge increase in *Any* common emotional annotations. Emotional diversity per term group increases, without no convergence of a common dominant emotion annotation.

E. Limitations

Lexicon acquisition is a complex task that includes a mixture of objective and subjective tasks. While annotation of emotions is more subjective, annotation of linguistic elements (such as stop words, emotion shift terms, intensifiers etc.) is purely objective. Our proposed workflow works well in

[9] https://pypi.python.org/pypi/pyenchant/

[10] https://www.nltk.org/

emotional annotation but could be improved with regards to intensifier, negator and stop word annotation.

Crowd diversity in the annotation and evaluation process is another factor. While crowd contributors might annotate a part of the corpus, domain experts will annotate the whole corpus. However, the uniformity of individual judgement is replaced with the diversity and mass of contributors [40].

Subcomponents of the lexicon acquisition could be experimented and improved upon on an individual basis. Lemmatisation could be used instead of stemming to group terms, spell check can include spelling recommendations, filtering could incorporate rewarding and penalties, evaluation process can include experts and so on.

The corpus may be limiting the term groups in the lexicon based on topic-specific submissions. Comparisons with existing lexicons, such as NRC[41] indicate an overlap of 40% terms. The rest of 60% terms in our lexicon are not present in NRC. Additionally, the lexicon could benefit from higher redundancy, as the mean number of annotations per term group is at 3.2.

V. Conclusions

We presented a multivalued lexicon acquisition process driven by the crowd. The resulting emotion association lexicon includes all the information obtained from crowdsourcing, is scalable, and presents a novel approach to evaluating subjective crowdsourcing tasks. The evaluation from the crowd is compared to the evaluation from domain experts. The comparison results provide a strong indication of the evaluating capabilities of the crowd.

Stemming reduces crowd costs and lexicon size. The multivalue approach of the lexicon, and the absence of aggregation or consolidation of annotations, highlight the subjective nature of the task and improve scalability. The peer evaluation of crowd contributions is almost identical to the expert evaluation of the contributions, with lower costs and faster responses. The obtained Likert-type scale evaluations can be used to determine terms that would benefit from further annotations, or signify terms that are considered of high quality, regardless of their received annotations and answer distribution.

The proposed lexicon creation workflow can be used as the acquisition process for a multitude of text and property association lexicons. We aim to explore personalised feedback systems that will provide actionable responses based on termaction association. Some of the topics we are also keen on exploring are political campaign polls with political stance and term association, or health issues and perceived severity. The common denominators of natural language applications are human intelligence and perception.

References

[1] Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language*, 28(1):20–37, 2014.

[2] A Adedamola, Abiodun Modupe, and O Dehinbo. Development and evaluation of a system for normalizing internet slangs in social media texts. *proeedings of WCECS*, 2015.

[3] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*, pages 30–38. Association for Computational Linguistics, 2011.

[4] Cecilia Ovesdotter Alm. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 107–112. Association for Computational Linguistics, 2011.

[5] Omar Alonso, Daniel E Rose, and Benjamin Stewart. Crowdsourcing for relevance evaluation. In *ACM SigIR Forum*, volume 42, pages 9–15. ACM, 2008.

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

[7] Muhammad Zubair Asghar. Detection and scoring of internet slangs for sentiment analysis using sentiwordnet. *Life Science Journal*, 11(9), 2014.

[8] Muhammad Zubair Asghar, Aurangzeb Khan, Shakeel Ahmad, and Fazal Masud Kundi. A review of feature extraction in sentiment analysis. *Journal of Basic and Applied Scientific Research*, 4(3):181–186, 2014.

[9] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204, 2010.

[10] Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. Sentiment analysis in the news. *arXiv preprint arXiv:1309.6202*, 2013.

[11] Anil Bandhakavi, Nirmalie Wiratunga, Stewart Massie, and Deepak Padmanabhan. Lexicon generation for emotion detection from text. *IEEE intelligent systems*, 32(1):102–108, 2017.

[12] Avinoam Baruch, Andrew May, and Dapeng Yu. The motivations, enablers and barriers for voluntary participation in an online crowdsourcing platform. *Computers in Human Behavior*, 64:923–931, 2016.

[13] Ria Mae Borromeo and Motomichi Toyama. Automatic vs. crowdsourced sentiment analysis. In *Proceedings of the 19th International Database Engineering & Applications Symposium*, pages 90–95. ACM, 2015.

[14] Daren C Brabham. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence*, 14(1):75–90, 2008.

[15] Michel Cabanac. What is emotion? *Behavioural processes*, 60(2):69–83, 2002.

[16] Chloé Cabot, Lina F Soualmia, Badisse Dahamna, and Stéfan J Darmoni. Sibm at clef ehealth evaluation lab 2016: Extracting concepts in french medical texts with ecmt and cimind. In *CLEF (Working Notes)*. CLEF, 2016.

[17] Erik Cambria, Bjorn Schuller, Bing Liu, Haixun Wang, and Catherine Havasi. Knowledge-based approaches to concept-level sentiment analysis. *IEEE Intelligent Systems*, 28(2):12–14, 2013.

[18] Jorge Carrillo-de Albornoz and Laura Plaza. An emotion-based model of negation, intensifiers, and modality for polarity and intensity classification. *Journal of the American Society for Information Science and Technology*, 64(8):1618–1633, 2013.

[19] Yoonjung Choi and Janyce Wiebe. +/-effectwordnet: Sense-level lexicon acquisition for opinion inference. In *EMNLP*, pages 1181–1191, 2014.

[20] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd international conference on computational linguistics: posters*, pages 241–249. Association for Computational Linguistics, 2010.

[21] Paul Ekman, E Richard Sorenson, Wallace V Friesen, et al. Pan-cultural elements in facial displays of emotion. *Science*, 164(3875):86–88, 1969.

[22] Maike Erdmann, Kazushi Ikeda, Hiromi Ishizaki, Gen Hattori, and Yasuhiro Takishima. Feature based sentiment analysis of tweets in multiple languages. In *International Conference on Web Information Systems Engineering*, pages 109–124. Springer, 2014.

[23] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI*, volume 7, pages 1606–1611, 2007.

[24] Manoochehr Ghiassi, James Skinner, and David Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications*, 40(16):6266–6282, 2013.

[25] Emitza Guzman and Walid Maalej. How do users like this feature? a fine grained sentiment analysis of app reviews. In *Requirements Engineering Conference (RE), 2014 IEEE 22nd International*, pages 153–162. IEEE, 2014.

[26] Mohammad Sadegh Hajmohammadi, Roliana Ibrahim, and Ali Selamat. Bi-view semi-supervised active learning for cross-lingual sentiment classification. *Information Processing & Management*, 50(5):718–732, 2014.

[27] Hussam Hamdan, Patrice Bellot, and Frederic Bechet. Sentiment lexicon-based features for sentiment analysis in short text. In *In Proceeding of the 16th International Conference on Intelligent Text Processing and Computational Linguistics*, 2015.

[28] Indukuri Hemalatha, GPS Varma, and A Govardhan. Automated sentiment analysis system using machine learning algorithms. *IJRCCT*, 3(3):300–303, 2014.

[29] Alexander Hogenboom, Daniella Bal, Flavius Frasincar, Malissa Bal, Franciska de Jong, and Uzay Kaymak. Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 703–710. ACM, 2013.

[30] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*, pages 607–618. ACM, 2013.

[31] Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 537–546. ACM, 2013.

[32] Clayton J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*, 2014.

[33] Hari Iyer, Mihir Gandhi, and Sindhu Nair. Sentiment analysis for visuals using natural language processing. *structural science*, 128(6), 2015.

[34] Aamera ZH Khan, Mohammad Atique, and VM Thakare. Combining lexicon-based and learning-based methods for twitter sentiment analysis. *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)*, page 89, 2015.

[35] Adam Kilgarriff and Christiane Fellbaum. Wordnet: An electronic lexical database, 2000.

[36] Svetlana Kiritchenko and Saif M Mohammad. The effect of negators, modals, and degree adverbs on sentiment composition. In *Proceedings of NAACL-HLT*, pages 43–52, 2016.

[37] Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762, 2014.

[38] Yue Lu, Malu Castellanos, Umeshwar Dayal, and ChengXiang Zhai. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th international conference on World wide web*, pages 347–356. ACM, 2011.

[39] Andrew Mao, Ece Kamar, Yiling Chen, Eric Horvitz, Megan E Schwamb, Chris J Lintott, and Arfon M Smith. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In *First AAAI conference on human computation and crowdsourcing*, 2013.

[40] M Lynne Markus. Toward a critical mass theory of interactive media universal access, interdependence and diffusion. *Communication research*, 14(5):491–511, 1987.

[41] Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*, 2013.

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

[43] Igor Mozetič, Miha Grčar, and Jasmina Smailović. Multilingual twitter sentiment classification: The role of human annotators. *PloS one*, 11(5):e0155036, 2016.

[44] Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. Semeval-2016 task 4: Sentiment analysis in twitter. *Proceedings of SemEval*, pages 1–18, 2016.

[45] Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77. ACM, 2003.

[46] Alvaro Ortigosa, José M Martín, and Rosa M Carro. Sentiment analysis in facebook and its application to e-learning. *Computers in Human Behavior*, 31:527–541, 2014.

[47] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[48] Robert Plutchik. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1(3-31):4, 1980.

[49] Robert Plutchik. The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001.

[50] Soujanya Poria, Erik Cambria, and Alexander F Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *EMNLP*, pages 2539–2544, 2015.

[51] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59, 2016.

[52] Soujanya Poria, Alexander Gelbukh, Amir Hussain, Newton Howard, Dipankar Das, and Sivaji Bandyopadhyay. Enhanced senticnet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, 28(2):31–38, 2013.

[53] Hamid Poursepanj, Josh Weissbock, and Diana Inkpen. uottawa: System description for semeval 2013 task 2 sentiment analysis in twitter. *Atlanta, Georgia, USA*, page 380, 2013.

[54] Yanghui Rao, Jingsheng Lei, Liu Wenyin, Qing Li, and Mingliang Chen. Building emotional dictionary for sentiment analysis of online news. *World Wide Web*, 17(4):723–742, 2014.

[55] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 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, 2008.

[56] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, et al. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

[57] Veselin Stoyanov and Claire Cardie. Topic identification for fine-grained opinion analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 817–824. Association for Computational Linguistics, 2008.

[58] Carlo Strapparava, Alessandro Valitutti, et al. Wordnet affect: an affective extension of wordnet. In *Lrec*. Citeseer, 2004.

[59] Duyu Tang, Furu Wei, Bing Qin, Ting Liu, and Ming Zhou. Coooolll: A deep learning system for twitter sentiment classification. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 208–212, 2014.

[60] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *ACL (1)*, pages 1555–1565, 2014.

[61] Cornelis J Van Rijsbergen, Stephen Edward Robertson, and Martin F Porter. *New models in probabilistic information retrieval*. British Library Research and Development Department, 1980.

[62] Casey Whitelaw, Navendu Garg, and Shlomo Argamon. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 625–631. ACM, 2005.

[63] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

[64] Jichang Zhao, Li Dong, Junjie Wu, and Ke Xu. Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1528–1531. ACM, 2012.

[65] Feng Zhou, Roger Jianxin Jiao, and Julie S Linsey. Latent customer needs elicitation by use case analogical reasoning from sentiment analysis of online product reviews. *Journal of Mechanical Design*, 137(7):071401, 2015.

[66] George K Zipf. Human behavior and the principle of least effort, 1950.