Demonstrating the Reliability of Self-Annotated Emotion Data

Anton Malko[1], **Cecile Paris**[1,2], **Andreas Duenser**[1], **Maria Kangas**[3], **Diego Mollá**[1,2], **Ross Sparks**[1], **Stephen Wan**[1]

[1]Data61, CSIRO, Australia
[2]Department of Computing, Macquarie University, Sydney, Australia
[3]Department of Psychology, Macquarie University, Sydney, Australia

anton.malko@csiro.au, cecile.paris@csiro.au, andreas.duenser@csiro.au, maria.kangas@mq.edu.au, diego.molla-aliod@mq.edu.au, ross.sparks@csiro.au, stephen.wan@csiro.au

Abstract

Vent is a specialised iOS/Android social media platform with the stated goal to encourage people to post about their feelings and explicitly label them. In this paper, we study a snapshot of more than 100 million messages obtained from the developers of Vent, together with the labels assigned by the authors of the messages. We establish the quality of the self-annotated data by conducting a qualitative analysis, a vocabulary-based analysis, and by training and testing an emotion classifier. We conclude that the self-annotated labels of our corpus are indeed indicative of the emotional contents expressed in the text and thus can support more detailed analyses of emotion expression on social media, such as emotion trajectories and factors influencing them.

1 Introduction

Social media platforms are being widely used by people to express their feelings. While some such platforms are generic in their purpose (e.g., Twitter), others have specific goals, such as connecting with people with similar health issues (e.g., PatientsLikeMe[1]). Vent[2] belongs to the latter class of platforms: its stated goal is to encourage people to express and share their feelings. Vent enables people to post messages expressing their own feelings and to react to posts from others. Interestingly, Vent requires people to label their posts with the emotion they feel at the time of posting. The platform thus provides us with an opportunity to study, at scale, how people express emotions, to what emotions they react, how emotions change over time, and what factors influence their trajectory.

Vent data is self-annotated for emotion, which is of particular interest to us. Studies on emotions in social media often derive labels from texts, either with the help of annotators, or using sentiment analysis techniques (see, for example, reviews of annotated datasets by Bostan and Klinger (2018); Mohammad (2020)). We note, however, that information that external observers (annotators or algorithms) can extract from a text may not be sufficient to reliably identify the affective state of the text's author at the time of posting. This could be because the texts are too short to provide enough context, are ambiguous, or require extra-textual context to interpret. Even when richer context is available, external observers may not necessarily assign a definitive affective label to a text. For example, psychological construction theory (Barrett, 2006) states that emotion labels are a result of *categorisation* of the current state of the organism, in the current context; consequently, the same episode may be categorised differently by the person who experiences it and by an outside observer. Given this, self-assigned affective labels may provide a more direct access to a person's emotional state than labels attributed after the fact.

Our ultimate goal is to study emotion trajectories (on social media) and the factors that affect them, potentially leading to the automatic identification of mental health issues. However, before we can employ data such as that provided in Vent to study emotion sharing and changes in emotions, we must establish whether the self-annotated labels are reasonable indicators of emotional states. This is because, even with self-assigned labels, there are concerns that may arise: for example, the label choice may be a byproduct of poor user interface design. Establishing that the labels are reasonable is thus our central aim in this paper. We conduct a multi-step analysis of the Vent data, showing that we can use this kind of data to study how people express their feelings and how people react to them.

[1]https://www.patientslikeme.com/
[2]https://www.vent.co/

Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology, pages 45–54

The rest of the paper is structured as follows. We begin with a short summary of related research in Section 2. This is followed by a description of the Vent platform and the data we have from it in Section 3. We describe the data selection steps in Section 4. We then present the analysis steps we have taken to ascertain that the labels adequately reflect the affective states expressed in the texts in Section 5. Section 6 concludes this paper and outlines future research directions.

2 Related Work

There is a growing number of datasets annotated with affect information. Many of these are annotated by experts or via crowdsourcing and fall out of the scope of our work. Instead, we refer the reader to the surveys by Bostan and Klinger (2018); Mohammad (2020).

To the best of our knowledge, self-annotated affective datasets are rare; the reviews by Bostan and Klinger (2018); Mohammad (2020) mention only one such dataset. ISEAR ("International Survey on Emotion Antecedents and Reactions") is a self-labelled affective dataset created by Scherer and Wallbott (1994). It was collected by administering a questionnaire, in which people were asked to describe recent experiences of one of the seven emotions (Anger, Fear, Joy, Sadness, Disgust, Shame, Guilt) and to answer questions about their physiological and psychological state during these emotion episodes. Overall, roughly 3,000 people from 37 countries completed the questionnaire, providing 7,666 textual descriptions. In comparison, our dataset contains considerably more data.

A more widely used approach to produce emotion annotation without using experts is to rely on distant supervision — for example, treating Twitter hashtags like `#happy` or `#sad` as self-assigned emotion labels. Examples of datasets constructed with distant supervision include those by Mohammad (2012); Roberts et al. (2012); Wang et al. (2012); Qadir and Riloff (2013); Mohammad and Kiritchenko (2015); Volkova and Bachrach (2016); Abdul-Mageed and Ungar (2017). Emotion classifiers using these datasets are reported to perform well: the best results thus far were produced by Abdul-Mageed and Ungar (2017), who used a Gated Recurrent Neural Network (GRNN) classifier on 1.6 million tweets labelled with emotions from Plutchik's categorisation (Plutchik, 1980) and reached an averaged F1-score[3] of 0.9568.

Lykousas et al. (2019) used web-scraping techniques to collect 33 million messages from the Vent platform, from around 1 million users with public profiles (meaning that anybody on the platform could see these posts). They presented a broad descriptive exploration of these data, along with an analysis of emotions in texts and user networks, but they did not investigate the quality of the annotations. In comparison, our dataset is directly provided via a 2019 data science partnership with Vent.

Our data includes all posts (anonymised for this research). Our goal here is to assess the alignment between affect in self-assigned affective labels and texts.

3 Vent and its Dataset

Vent advertises itself as a platform to "Express your feelings and connect with people who care". Vent is thus specifically geared towards sharing one's emotions, unlike Twitter or Facebook, which support many other activities. This makes Vent particularly interesting for investigating emotion expression on social media. Users (*venters*) register anonymously, with only an email address. Once registered, they can create short text messages (*vents*), read messages by other venters and react to them, using comments or *interactions* (short predefined reactions, for example, "HUG", "LOL", or an emoji).

Vent's creators have given us access to the data from the platform over a 5-year period, from the late 2013 until the end of 2019, as part of a collaborative project to study mental health.[4] Overall, the raw dataset contains over 107 million vents, from close to 1.5 million users, including both public and private posts, along with additional types of information, namely comments, interactions, follower/followee links and the information on discussion groups. Due to ethical and privacy concerns, the dataset is not publicly available.

Vent's labels[5] are arranged in a two-level hier-

[3] A classification performance metric, which takes into account both the classifier's accuracy on the target class (Recall), and its ability to avoid classifying non-target examples as target (Precision). It is defined as a harmonic mean of Precision and Recall; its worst value is 0 and its best value is 1 (Chicco and Jurman, 2020).

[4] The project was approved by the CSIRO ethics committee; reference number 165/19.

[5] For clarity, we will use different fonts to refer to Vent's label categories (e.g., `Sadness`) and real affective states (e.g., sadness).

archy. At the top level, there are 85 emotion categories, which we can categorise into these 5 groups:

Affective states. This group contains the following 9 categories: `Affection`, `Anger`, `Creativity`, `Fear`, `Feelings`, `Happiness`, `Positivity`, `Sadness`, `Surprise`.

Dates. There are 46 categories linked to dates and seasonal events, such as `Autumn`, `Ramadan`, `Paralympics`, etc.

Groups of people. There are 13 categories in this group, e.g., `Women HM`, `Pride'18`, etc.

Character/Role/Imaginary content. This group contains 7 categories related to fictional and imaginary topics such as `Vampire`, `Star Wars`.

Miscellanea. There are 10 categories of miscellaneous nature, e.g., `Candy`, `Gaming`.

The nine categories related to affective states are always available to the users. All other categories generally have to be paid for individually, although they can become temporarily available for free on special occasions (e.g., on Halloween).[6]

At the second level of the hierarchy, there are 1,187 labels. Figure 1 shows examples of labels within a subset of the 9 always available categories.

When users want to create a message, they first go through a labelling interface: all labels from a given category are presented on a single screen, and swiping the screen to the left or right switches between label categories. The name of the current category is *not* shown to the users by default and is only indicated by the background colour of the screen. It becomes visible if one taps on the scrolling control.

In the current version of Vent, when users create a new vent, their label choice screen starts with the label category from their most recent vent. This might introduce biases to the data: for example, users may just proceed with the first choice they see (e.g., if they need to share some intense emotion experience and accurate labelling is not important to them at the moment). We also note that people *have to* select a single label. Finally, the inventory of labels is pre-defined. In some situations, this may cause people to choose a label that does not exactly match their current dominant state.

[6]This has changed in the most recent versions of Vent: currently, one has to pay a monthly subscription fee to unlock all additional label categories.

4 Data selection

For our analyses, we restrict our data to the vents that correspond to the following six high-level categories, which we call "core categories": `Affection`, `Anger`, `Fear`, `Happiness`, `Sadness`, and `Surprise`. These labels are always available to the users. Importantly, out of all Vent's categories, they are most easily interpretable in terms of affective states. Many psychological accounts of human emotion repertoire include some or all of these categories (see, e.g., Table 1 of Ortony and Turner's (1990) publication); and they map one-to-one onto Shaver et al.'s (1987) classification. Vents with these labels account for 45.4% of the total number of vents.

In addition, we exclude the following categories of users:

1. **Official Vent account.** Vent has an official account, which consists mostly of a) questionnaires about experiences on Vent; b) technical information (e.g., planned maintenance) and c) discussion of possible/existing label categories.

2. **Robots.** The following heuristic was used: a user is a robot if (1) they created at least 100 messages within a day, (2) they posted vents on no more than 10 distinct dates, and (3) at least 99% of the vents were posted within a single day. Using this rule, we discovered 258 users, who created 187,063 messages. A manual analysis suggested that our heuristic is satisfactory: only 1 of 30 randomly selected users in this subset was not a robot. One additional robot account with 10,219 vents not satisfying the heuristic criterion was further excluded during manual exploration.

3. **Users with fewer than 20 vents.** The purpose of this filter was to ensure that the users we include have at least some experience in using the app.

The resulting dataset contains 45,194,018 vents from 372,662 users. It is used for the qualitative analysis in Section 5.1.

For the more detailed automated analyses in Sections 5.2–5.4, we further subset these data in the following way. Most categories contain labels which

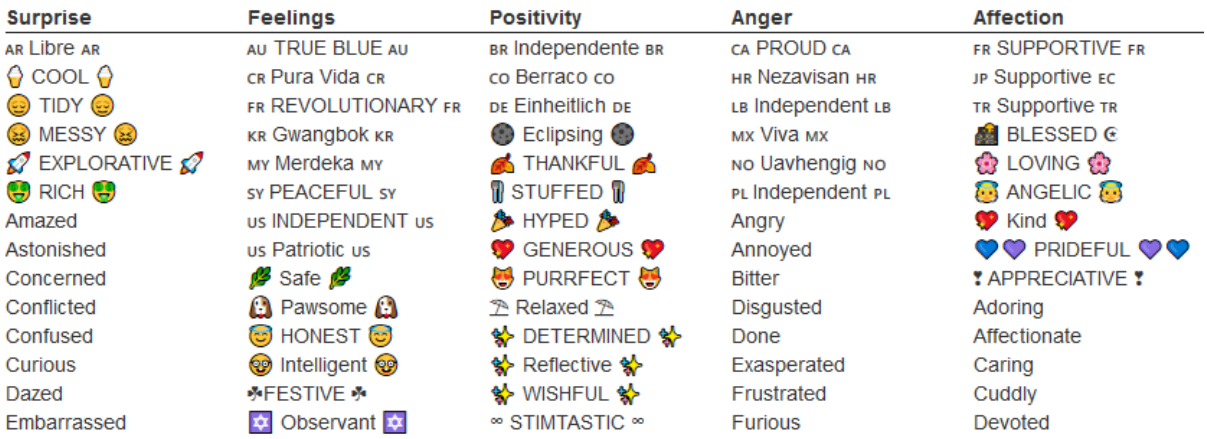

Figure 1: A selection of label categories and labels.

are less clearly connected to affect or only used rarely (e.g., "Independent" or "Viva" in `Anger` in Figure 1) — we exclude them from consideration. Next, we sample 1.8 million vents per core label category, filtering out (a) vents only containing words "null", "test" or "testing"; (b) tag memes. Tag memes are explained in Section 5.1 and are identified with a regular expression.[7] Finally, we exclude non-English vents, as identified by the `langid`[8] tool, which removes approximately 7% of the messages. The resulting subset contains 1.5–1.6 million messages per label category; we will refer to it as "the reduced dataset".

5 Assessing the alignment of the labels and the texts

To ascertain the alignment between the text and the labels, we conduct the following analyses:

1. A qualitative analysis, conducted manually on a subset of the data in order to identify potentially non-affective uses of the labels;

2. A vocabulary-based analysis, in which we gather statistics on the presence of emotionally loaded words in the vents using word-emotion associations;

3. An emoji-based analysis, in which we examine the top 10 emojis in each label category of interest; and

4. A text-to-label machine learning classifier analysis, in which we train a BERT model to establish whether textual information beyond simple keywords helps to differentiate between individual label categories.

We use these four methods to establish that the self-annotated labels do indeed reflect emotional state. The methods are complementary. The qualitative analysis attempts to capture idiosyncratic uses of labels, which may be hard to anticipate and thus hard to analyse automatically. The vocabulary-based and emoji-based analyses establish whether individual emotion-loaded tokens in the texts are congruent with the labels. Finally, the classification approach allows the exploration of the connection between entire texts and their labels, capitalising on context beyond individual tokens. These analyses are described below.

5.1 Qualitative analysis

During an initial data exploration, we found the following cases of non-affective uses of the labels:

1. **Vents with "default" labels**. Some people choose default labels for their vents, occasionally stating reasons for doing this: for example, liking the colour of a specific category or being too lazy to chose a label for every vent.

2. **Vents from bio accounts**. Vent allows users to add biographical information to their accounts; however, some users create separate dedicated accounts just to post messages containing such information. Posts in these accounts include not only demographic facts, but also topics of interest, and guidelines for followers (describing who should or should not follow).

[7] `(.*tagged by.*) | (.*i tag.*) | (.*tagging.*)`

[8] https://github.com/saffsd/langid.py

3. **Tag memes**: We observed the occurrence of user-generated questionnaires on a wide variety of topics (e.g., "What kind of vent user are you?", "common fears"). Vent users refer to them as "tag memes". Such questionnaires often follow a specific template, so we could identify them based on a regular expression. A manual analysis of 100 messages identified using the regular expression we employed showed that 18 of them were not tag memes.

To assess the relative presence of the above non-affective uses of the labels, we inspected 1,000 randomly selected vents from the dataset (after applying the filters described in Section 4). The sample did not contain instances of people mentioning default emotions. The sample contained 4 tag memes (0.4% of the sample), and only 1 vent from a bio account. We therefore conclude that clearly non-affective uses of labels are rare.

5.2 Vocabulary-based analysis

After performing the qualitative analysis, we consider emotionally loaded words present in the texts. The data used for this analysis is a sample of 1.5 million vents per category from the reduced dataset, to have a balanced distribution across categories.

The emotionally loaded words are obtained from the the NRC Emotion Lexicon (henceforth, EmoLex) (Mohammad and Turney, 2012). EmoLex is one of the largest emotion lexicons. It contains 14,182 words and indicates whether they are associated with one of 10 affective states: Plutchik's eight (Plutchik, 1980), plus "positive" and "negative". Each word can be associated with any number of affective states.

For this analysis, we only consider EmoLex words associated with at least one specific emotion, excluding words which only have generic associations with positive and/or negative affect. This results in 4,463 unique words out of 14,182 and 8,265 word-affect association pairs. Around 70% of the vents have words from this set.

Table 1 shows the lexicon coverage per label category. Within all label categories, except `Surprise`, words related to a corresponding emotion[9] are found in the largest proportion of vents. For example, if we consider vents labelled with `Anger` (second column of Table 1), EmoLex words related to anger are found in the largest proportion of these vents. Such associations also hold at the more general level of emotional valence (positive vs. negative affect): within a given label category, EmoLex words associated with emotions of matching valence are generally found in a larger proportion of vents: e.g., within `Sadness`, more vents contain words related to anger, fear and sadness than to anticipation, joy and trust.

Table 1: Percentage of vents having at least one word associated with a given emotion. 'Any' – proportion of vents with at least one word associated with any emotion. Af – `Affection`, An – `Anger`, Fe – `Fear`, Ha – `Happiness`, Sa – `Sadness`, Su – `Surprise`. Maximum values in each column (excluding the 'Any' row) are highlighted in bold.

		Vent label category					
		Af	An	Fe	Ha	Sa	Su
EmoLex emotion category	anger	24	**42**	34	23	34	27
	anticipation	37	33	38	40	32	33
	disgust	21	38	30	21	30	24
	fear	23	38	**39**	23	36	27
	joy	**47**	30	30	**42**	30	32
	sadness	25	40	38	23	**42**	28
	surprise	24	21	22	24	21	20
	trust	39	35	35	40	32	**35**
	any	69	72	71	69	68	66

Conversely, if we examine what Vent category has the largest percentage of words from a determined EmoLex emotion category (by analysing Table 1 row by row, instead of column by column), we observe that closely related categories are most likely. For example, if we know that a vent has sadness-related words, it is most likely to be labelled with `Sadness`. This pattern holds in virtually all cases when there exists a one-to-one mapping from an emotion to a Vent category, with only two exceptions: EmoLex words associated with joy are most likely to be found in `Affection` vents, and EmoLex words associated with surprise are most likely to be found in `Affection` and `Happiness` vents. A similar pattern is observed for emotion valence: for example, words associated with anger are most likely to be found in vents labelled with any category with negative valence: `Anger`, `Fear`, `Sadness`. These results suggest

[9] Vent category of `Affection` does not have a corresponding emotion in EmoLex, but arguably, joy is the closest option. Plutchik considered love to be a combination of joy and trust (e.g., see (Plutchik, 1980, p.21); "trust" is called "acceptance" in the reference), and interestingly, high proportion of `Affection` vents have words related to these emotions.

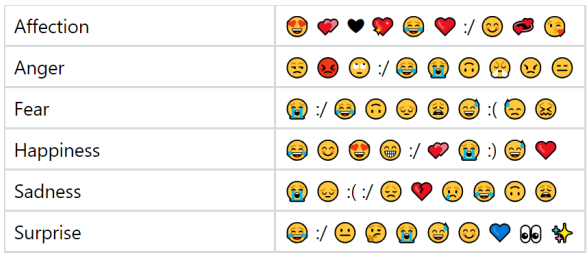

Figure 2: Top 10 emojis per category. Emojis are ordered from most to least used.

that when people use words associated with a given emotion, they are more likely to choose the corresponding Vent label.

5.3 Emoji based analysis

The previous section showed that the data was consistent with the EmoLex resource. We perform a similar analysis with emojis, which are not included in EmoLex, checking to see that these distant supervision labels are generally consistent with the self-annotated labels. We carry out a separate analysis of the most used emojis, using the same dataset of 1.5 million vents per label category. Emojis were identified using the `emoji`[10] and `emot`[11] Python libraries.

Figure 2 shows that the use of emojis is congruent with the category. For example, the top 10 emojis in `Affection` contain more hearts than any other category; and emojis indicating angry faces only appear in the top 10 list for `Anger`. We can observe the same at the level of affect valence as well. For example, the "`: (`" emoticon does not appear in the top 10 list for `Affection` and `Happiness`; hearts do not appear in the top 10 list for `Anger`, `Fear` and `Sadness` (with the exception of the *broken* heart in `Sadness`).

This analysis of the use of emojis per Vent category is consistent with the vocabulary-based analysis of the previous section.

5.4 Emotion classification

Our final analysis to assess the alignment of the labels and the texts has been conducted by training a neural emotion classifier with the Vent data, and observing the results on a separate test data, also drawn from the Vent data. The rationale is that, if the classifier can identify the labels, then these labels are used in a consistent way. Of course this does not indicate *per se* that the self-annotated labels are correct, because there might have been a bias in the labelling process which has been captured as a pattern picked by the classifier. But combined with the analysis described in the previous sections, good classification results would give additional evidence for the validity of the self-annotated data.

[10] https://github.com/carpedm20/emoji/
[11] https://github.com/NeelShah18/emot

Table 2: EmoLex-based models. F1-score by class: mean value (stddev) across the five runs

Label category	Precision	Recall	F1
`Affection`	–	–	–
`Anger`	0.26 (0.002)	0.16 (0.002)	0.20 (0.002)
`Fear`	0.25 (0.004)	0.15 (0.002)	0.19 (0.003)
`Happiness`	0.29 (0.002)	0.19 (0.002)	0.23 (0.002)
`Sadness`	0.27 (0.003)	0.19 (0.003)	0.22 (0.003)
`Surprise`	0.22 (0.003)	0.08 (0.001)	0.12 (0.002)

For the classification data, we create 5 random subsets with 500,000 + 50,000 + 50,000 vents (train-dev-test) per category, each time sampling from the reduced dataset. All texts are lowercased.

We use two simple classifiers as baselines. In the first one, labels are simply chosen at random from Vent's core categories. This classifier produces Precision of 0.17, Recall of 0.17 and F1-score of 0.17 for all classes. The second classifier is based on EmoLex. For each vent in our sample, we predict the EmoLex emotion associated with the largest number of words in this vent. Ties (including cases where vents contained no words from EmoLex) are broken at random. As Vent's `Affection` does not map directly onto EmoLex emotions, we exclude it from consideration in this particular analysis. The classification results generally improve over the random baseline, but the gains are small: the macro F1-score ranged from 0.189 to 0.192, with a mean of 0.190 and a standard deviation of 0.001. The F1 scores by class averaged across all five runs are given in Table 2.

Finally, we use a BERT-based model (De-

Table 3: BERT-based models. F1-score by class: mean value (stddev) across the five runs.

Label category	Precision	Recall	F1
`Affection`	0.62 (0.005)	0.65 (0.005)	0.63 (0.005)
`Anger`	0.57 (0.005)	0.57 (0.004)	0.57 (0.000)
`Fear`	0.54 (0.005)	0.49 (0.005)	0.52 (0.005)
`Happiness`	0.58 (0.000)	0.59 (0.004)	0.58 (0.005)
`Sadness`	0.54 (0.005)	0.60 (0.000)	0.56 (0.004)
`Surprise`	0.52 (0.004)	0.47 (0.004)	0.49 (0.004)

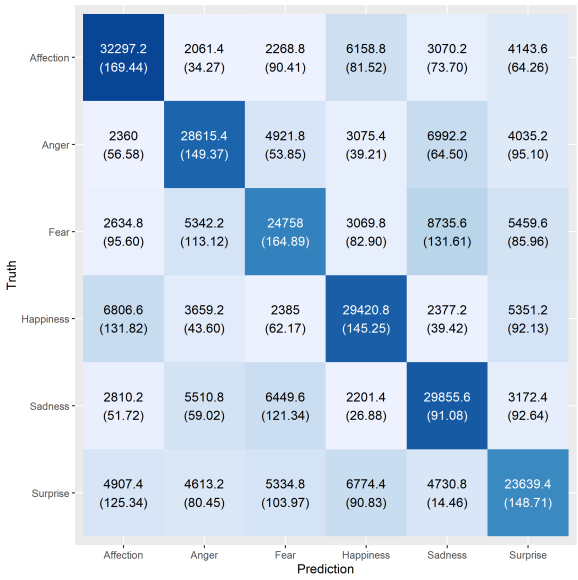

Figure 3: Confusion matrix for BERT model's predictions. Numbers correspond to mean values (stddev) across the five runs.

vlin et al., 2019).[12] The model's standard lexicon is manually augmented with emojis, using `emoji2vec` pre-trained embeddings (Eisner et al., 2016). We use the following hyperparameters. The maximum sequence length for the BERT tokeniser is set at 128. The learning rate is $3 \cdot 10^{-5}$. The batch size is 512 (spread over 4 GPUs). The number of epochs is 2, with checkpoints every 150 batches. The best checkpoint (as measured by macro F1) is saved.

We train a separate model on each random subset. Macro F1 score ranges from 0.560 to 0.562, with a mean of 0.561 and a standard deviation of 0.001. Table 3 shows F1-score by class, and Figure 3 shows the confusion matrix for the model's predictions; in both cases the values are averaged across the five runs.

The BERT-based classifier has improved performance, indicating that context over and above emotionally loaded keywords contains considerable amount of information benefiting classification. With respect to the alignment between labels and texts, the results are consistent with the results of the vocabulary-based and emoji-based analyses (Figure 3). The correct label is predicted most frequently. Incorrectly predicting a label referring to the emotion of similar valence is more likely than predicting a label of the opposite valence: e.g., when the true label is `Happiness`, `Affection` is a more likely incorrect prediction than `Anger`, `Fear` or `Sadness`. As before, the category of `Surprise` appears to be less clearly connected with the texts properties: the classifier made the biggest number of mistakes on it, and these mistakes were relatively evenly spread across the other 5 categories.

To better understand the classifier's performance, we visually inspect 60 random sentences (10 per label category) in which the classifier made a wrong prediction. Given that the variability between the models in the five runs is small, we only examine predictions from a single model with the best macro F1 score. Table 4 shows the results. As recommended by Benton et al. (2017), all specific examples are rephrased to protect users privacy. In the majority of the vents (45), the label assigned by the classifier is consistent with the text. Common reasons for the mistakes include lack of context which would allow to clearly differentiate between several possible affective states (e.g., `Affection` and `Happiness`, or `Anger` and `Sadness`); multiple emotions clearly expressed in the text (in some cases the classifier did capture one of the emotions, while the label reflected another). In a minority of cases, it is not immediately clear whether the labels fit the text (8 cases). In two such cases, the orthography is quite severely affected. In four cases, the Vent label hierarchy is to blame: the lower level label matched the sentence, but the category it be-

[12] `bert-base-uncased` from the HuggingFace Transformers library (Wolf et al., 2019).

Table 4: Analysis of 60 random examples in which there is a mismatch between the gold label existing in Vent and the automated label assigned by the classifier. "No context" — not enough context to assign a label, given just text. "Both" — both the gold and the automated label fit the sentence, and (a) "Both conceivable" — it is hard to choose between them; (b) "Gold better" — the gold label appears a better fit; (c) "Automated better" — the automated label appears to be a better fit. "Gold only" — only the gold label fits. "Automated only" — only the automated label fits. "Neither" — neither the gold nor the automated labels fit. Examples are accompanied by the gold label (in bold) and the automated label.

Type	**Count**	**Example**
No context	3	(1) "*Ahaa*" (**Anger**; Affection)
Both	45	
Both conceivable	26	(2) "*It seems I am always the problem*" (**Anger**; Sadness)
Gold better	10	(3) "*Nowadays movies are very strange*" (**Surprise**, Happiness)
Automated better	9	(4) "*Why can't I fall asleep. It's always this way, I want to sleep and not be stressed. Everything is going to be even worse tomorrow. I just wanna f***ing sleep... [several more similar sentences]*" (**Fear**; Anger)
Gold only	4	(5) "*This crazy woman told me to stop watching animes and study instead. My animes have more culture than you.*" (**Anger**; Happiness)
Automated only	2	(6) "*I hate friends who do what you ask them not to. If I tell not to look at me, f***ing don't. F***ING LISTEN TO ME*" (**Fear**; Anger)
Neither	6	(7) "*Can't wait until the evening, I do need some time for myself*" (**Fear**; Happiness)

longed to did not. One example is the vent "*I am leaving tomorrow, this is sad, but also a relief, as I am tired and want to be home.*" — the lower level label is "Stressed", which is congruent with the text; however this label falls under `Fear` category, which is a worse fit for the message.

The model performance, and consistency with the vocabulary and emoji analysis performed in Sections 5.2–5.3, gives further evidence that the affective information contained in vents is congruent with the assigned labels.

6 Conclusions

In this paper, we have presented an analysis of the quality of self-annotated emotion data from the Vent platform, which is specifically focused on emotion sharing. Our results suggest that self-assigned labels in Vent have a reasonable degree of connection to the affective states expressed in the texts. A qualitative analysis of the vents and their labels indicates that labels which are not meant to communicate affect are rare. A vocabulary-based analysis based on EmoLex shows that Vent labels align with affect polarity of the texts, and that words associated with a certain EmoLex emotion are most frequently encountered in vents in the corresponding Vent category. The top 10 emojis in each category are consistent with the category label. Finally, a BERT classification model can predict correct labels most often, and the classification mistakes often preserve emotion valence. Overall, we conclude that self-assigned labels produced in a non-controlled naturalistic setting can be used as a reasonably accurate representation of the author's affective state, and thus can support more complex analyses of emotions in social media.

Our analyses focused on the assumption that each text conveys one dominant emotion which may or may not be congruent with the assigned label. We adopted this approach as a first step, allowing us to explore simple models matching the structure of the data (one message – one label). This is an oversimplification, as suggested by examples such as (4) in Table 4 or the earlier example about going home (*"this is sad, but also a relief"*). Several emotions may be expressed in a

single text, either because the emotional state of the author evolved during the writing of the message, or because the author had mixed emotions (e.g., Larsen and McGraw (2014)). As Vent only allows one label per message, the presence of vents containing mixed emotions could lower the observed alignment between the labels and the texts.[13] Thus, understanding whether and how mixed emotions are expressed in naturalistic data such as those from Vent would be important in this line of research, and we may explore it in our future work.

One particular research direction we are currently exploring is tracking the changes in reported emotion over time, the factors influencing these changes, and the connection of these properties with mental health well-being.

Acknowledgements

We would like to thank Dean Serroni and Albert Jou from Vent for granting us access to the data and for providing additional information about the data and the app. We would also like to thank the two anonymous reviewers for providing feedback on an earlier version of this paper.

References

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

Lisa Feldman Barrett. 2006. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, 10(1):20–46.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[13]For example, an anonymous reviewer raised a possibility that the BERT classifier might have showed lower results on `Fear` and `Surprise`, because these emotions are more commonly associated with mixed emotional experiences.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54, Austin, TX, USA. Association for Computational Linguistics.

Jeff T. Larsen and A. Peter McGraw. 2014. The case for mixed emotions. *Social and Personality Psychology Compass*, 8(6):263–274.

Nikolaos Lykousas, Costantinos Patsakis, Andreas Kaltenbrunner, and Vicenç Gómez. 2019. Sharing emotions at scale: The vent dataset. In *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)*.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Saif M. Mohammad. 2020. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. *Emotion measurement*.

Saif M Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Saif M. Mohammad and Peter D. Turney. 2012. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Andrew Ortony and Terence J. Turner. 1990. What's basic about basic emotions? *Psychological Review*, 97(3):315–331.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of Emotion*, pages 3–33. Elsevier.

Ashequl Qadir and Ellen Riloff. 2013. Bootstrapped learning of emotion hashtags #hashtags4you. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media*

Analysis, pages 2–11, Atlanta, Georgia. Association for Computational Linguistics.

Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. 2012. EmpaTweet: Annotating and detecting emotions on Twitter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3806–3813, Istanbul, Turkey. European Language Resources Association (ELRA).

Klaus R. Scherer and Harald G. Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66(2):310–328.

Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'Connor. 1987. Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52(6):1061–1086.

Svitlana Volkova and Yoram Bachrach. 2016. Inferring perceived demographics from user emotional tone and user-environment emotional contrast. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1567–1578, Berlin, Germany. Association for Computational Linguistics.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing twitter" big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592. IEEE.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv preprint*.