Exploring Reliability of Gold Labels for Emotion Detection in Twitter

Sanja Štajner
Symanto Research
Nuremberg, Germany
sanja.stajner@symanto.com

Abstract

Emotion detection from social media posts has attracted noticeable attention from natural language processing (NLP) community in recent years. The ways for obtaining gold labels for training and testing of the systems for automatic emotion detection differ significantly from one study to another, and pose the question of reliability of gold labels and obtained classification results. This study systematically explores several ways for obtaining gold labels for Ekman's emotion model on Twitter data and the influence of the chosen strategy on the manual classification results.

1 Introduction

Interest for automatic emotion detection has been gaining popularity in the last ten years (Acheampong et al., 2020, Figures 3 and 4). The span of its applications ranges from empathetic chatbots and virtual agents (Paiva et al., 2017; Rashkin et al., 2019; Lin et al., 2019b; Shin et al., 2019; Lin et al., 2019a; Ma et al., 2020) to social media and public opinion analysis (e.g. (Anstead and O'Loughlin, 2014; Wu et al., 2020; Loureiro and Alló, 2020)). Nevertheless, the task proved challenging, especially when attempted at purely textual utterances as opposed to the multi-modal ones (Poria et al., 2019), probably due to missing visual and audio cues (Acheampong et al., 2020).

Previous studies reported some of the challenges in automatic emotion detection from texts: different perspectives one may take (Buechel and Hahn, 2017b; Alm et al., 2005), missing context (Öhman et al., 2020; Mohammad, 2012), non-literal meaning (Mohammad, 2012), high subjectivity of the task and low inter-annotator agreement even among trained annotators (Alm et al., 2005; Schuff et al., 2017). For example, the utterance *"Italy defeats France in World Cup Final"* (Katz et al., 2007) is most probably neutral from the writer's (journalist's) perspective, while evoking strong and probably opposite emotions among Italian and French readers (Buechel and Hahn, 2017b). The utterance *"Time for shopping"* might be neutral, or express/evoke various emotions (e.g. *joy*, *anger*, *fear*) depending on the writer's/reader's associations and personal experiences with shopping.

The field of emotion detection from text, similar to many other areas of natural language processing, suffers from the absence of standards for human annotation, and systematic investigations of how different strategies for obtaining gold labels influence classification results. Notable exceptions are the studies by Mohammad and Turney (2013) and Buechel and Hahn (2017b). Mohammad and Turney (2013) found that asking annotators which emotion is the word *associated with* yields higher inter-annotator agreement than asking them which emotion the word *evokes*. This result indicated that annotating emotions from text's perspective is less subjective than annotating them from reader's perspective. Motivated by those results, Buechel and Hahn (2017b) investigated the influence of perspective on the inter-annotator agreement in emotion annotation at a sentence level.

A recent study by Northcutt et al. (2021) demonstrated that incomplete or suboptimal gold labels in benchmark datasets can steer research efforts in wrong direction as they reward systems that comply with such suboptimal labels. Obtaining the correct gold labels for emotion detection from texts should thus be of the utmost importance for the field.

This study's contributions towards that goal are:

- An overview of previous efforts in human annotation of emotions in texts (Section 2).
- Single-label human annotation of Ekman's emotions in English tweets by six trained annotators (Section 3).

Proceedings of Recent Advances in Natural Language Processing, pages 1350–1359
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_151

Study	#annotators		Gold	#emotions	Labelling	Perspective	Genre
	Per instance	Total					
(Demszky et al., 2020)	3 or 5	82	$>$ 1 annotator	27+1	multi	writer	Reddit
(Bostan et al., 2020)	5	310	$>$ 1 annotator	15+1	single	text	Headlines
(Öhman et al., 2020)	$\leq$3	108	$>$ 1 annotator	8+1	multi	speaker	Subtitles
(Poria et al., 2019)	5	?	majority	6+1	single	speaker	Dialog
(Hsu et al., 2018)	5	?	majority*	6+1	single	speaker	Dialog
(Schuff et al., 2017)	3–6	6	various	8	multi	?	Twitter
(Mohammad et al., 2015)	3+	$\approx$ 3000	$>$ half	19+1	single	text	Twitter
(Brynielsson et al., 2014)	3	3	majority	3+1	single	writer	Twitter
(Neviarouskaya et al., 2010)	3	3	$\geq$2 agree	14	single	?	Various
(Neviarouskaya et al., 2009)	3	3	$\geq$2 agree	9+1	single	?	Blogs
(Strapparava and Mihalcea, 2007)	6	6	?	6	multi	reader	Headlines
(Aman and Szpakowicz, 2007)	2	4	both agree	6+2	single	text	Blogs
(Alm et al., 2005)	2-3	3	majority	6+1	single	text	Children

Table 1: Annotation procedures used in previous studies ("?" signifies that the particular aspect was not specified in the paper, "+1" in the *#emotions* column signifies the additional class for "other" or "no emotion").

- Detailed analysis of the collected human annotations and their comparison to the automatically assigned labels that are the current standard for obtaining gold labels on Twitter data (Section 4.1).

- Systematical investigation of several strategies for obtaining gold labels from manual annotations, and their influence on the reported manual classification results (Section 4.2).

2 Related Work

Several recent surveys (Acheampong et al., 2020; Alswaidan and Menai, 2020) and studies (Öhman et al., 2020; Bostan et al., 2020; Bostan and Klinger, 2018; Schuff et al., 2017) list previous work on emotion detection from texts and emphasise their differences in type of emotion taxonomy, task (single-label or multi-label), size of the dataset, text genre, granularity, topics, system architectures, and best results obtained with systems for automatic detection of emotions in texts.

However, none of the studies focussed on assessing the quality of benchmark datasets, or the influence of methods used for obtaining gold labels on the results of systems for automatic emotion detection from texts.

Drawing conclusions about influence of strategies for obtaining gold labels on the classification results by systematic exploration of the previous studies is not possible due to different text genres, number and type of annotators (trained vs. crowdsourced), annotation type (single-label vs. multi-label), granularity of annotations (word or sentence level, with or without surrounding context), emotion taxonomies, and the perspective taken. Table 1 presents annotation strategies used in some of the previous studies. For instance, in a multi-labelling task with 27 emotions (+ neutral) where each Reddit comment was annotated by three or five annotators out of a total of 82 crowdsourced annotators, the Cohen's kappa (Cohen, 1960) was calculated by aggregating all pairs of annotations per instance and emotion (Demszky et al., 2020). In a single-labelling task with 15 emotions (+ no emotion) where each news headline was annotated by five out of 310 annotators, in contrast, the authors report the Fleiss' kappa (Fleiss, 1971) as a measure of inter-annotator agreement (Bostan et al., 2020). In the XED dataset of movie subtitles, annotated with 8 emotions (+ neutral) in a multi-labelling task, some instances were annotated with fewer than three annotators (some even one annotator only), and the Cohen's kappa was calculated for gold labels in the parallel dataset of movie subtitles in English and Finish (Öhman et al., 2020).

Some studies went beyond the "simple" emotion annotation, by requesting from annotators to annotate the intensity of the emotions, e.g. (Strapparava and Mihalcea, 2007; Aman and Szpakowicz, 2007; Buechel and Hahn, 2017a), the triples of *who experiences which emotion and why* (Kim and Klinger, 2018). In a recent study, Bostan et al. (2020) conducted the annotation of emotions, cues, intensities, experiencers, causes, targets, and reader's emotions on news headlines.

A commonly used strategy for obtaining large training datasets for emotion detection in texts is by automatically labelling tweets that contain hashtags that explicitly mention emotions or predefined emotion keywords. Wang et al. (2012) used 131 emotion hashtags as keywords for collecting 5 million tweets in seven emotion categories (*joy*, *sadness*, *anger*, *love*, *thankfulness*, *surprise*). They explored several different strategies for obtaining gold labels based on hashtags and found that most accurate gold labels are obtained if the keyword hashtag appears at the end of the tweet; keyword hashtags appearing anywhere else in the tweet were not found to be that relevant. Shahraki and Zaïane (2017) automatically annotated tweets with nine emotions (*anger*, *fear*, *joy*, *love*, *sadness*, *surprise*, *thankfulness*, *disgust* and *guilt*) based on 15 explicit hashtags appearing in them, resulting in Clean Balanced Emotional Dataset (CBET) with 27,000 annotated tweets (3,000 per each emotion). Mohammad (2012) compiled a corpus of 21,051 tweets which contained one of the six Ekman's emotions (*anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*) as the last hashtag, and suggested to use it as an automatically assigned label that corresponds to the emotion experienced by the writer. One of the findings of that study was that such large, automatically-labelled training dataset can be used for emotion detection in other domains and text genres.

3 Methodology

3.1 Dataset

The dataset used in this study is a subset of TEC dataset (Mohammad, 2012), available through the Unify dataset (Bostan and Klinger, 2018). It consists of 35 randomly selected English instances for each of the six Ekman's emotions (*anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*) that contain at least six tokens. To select 35 instances of each of the six emotions, the gold labels of the original dataset were used, i.e. the automatically-assigned gold labels based on the last hashtag in the post, as described in the previous section, e.g. *"We are fighting for the 99% that have been left behind. #OWS #anger"* (Mohammad, 2012).

3.2 Annotation Procedure

Annotators. Each of the 210 tweets (35 per each emotion class) was annotated by six trained, paid annotators. Three annotators were male and three female. All annotators had at least a bachelor degree. Two of the annotators (one male and one female) were native speakers of English (UK); the other four annotators were proficient in English, and use it in their everyday work.

Guidelines. The annotators were instructed to choose, for each tweet, from a drop-down menu, one of the seven possible labels (ANGER, DISGUST, FEAR, JOY, SADNESS, SURPRISE, NEUTRAL), which best represents the emotion of the writer of the post, as the automatically assigned gold labels were expected to represent writers' emotions as well (Mohammad, 2012).

Evaluation. The inter-annotator agreement, either between the gold label and one of the annotators, or between two annotators, was calculated in two ways: as accuracy (the percentage of cases in which the given labels match); and Cohen's κ (Cohen, 1960).[1]

3.3 Experiments

Two sets of experiments were conducted. In the first set of experiments, the results of the human annotation experiment were analysed and compared with the automatically obtained gold labels. The goals of this set of experiments were: (1) to investigate reliability of automatically obtained gold labels; and (2) to estimate complexity of the task for trained human annotators and find the main causes of their disagreements.

In the second set of experiments, several strategies for obtaining gold labels from human annotations, and their influence on the manual classification results were explored, given that previous studies used various strategies for obtaining gold labels from human annotations (Table 1, Section 2). The explored strategies for obtaining gold labels were: (1) based on the last emotion-hashtag (no human annotation required); (2) based on the annotations of just one trained annotator; (3) based on the majority label obtained from three human annotations; and (4) based on the majority label obtained from five human annotations.

4 Results and Discussion

The results of the two sets of experiments are presented and discussed in two separate subsections.

4.1 Annotation Analysis

The pairwise inter-annotator agreements for each pair of annotators, and for each annotator and the

[1] For calculation of Cohen's κ, we used the implementation in sklearn library for python: `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html`.

Annotation pair	Cohen's κ			Agreement		
	min	max	avg	min	max	avg
Two annotators	0.33	0.55	0.41	43.3%	63.3%	50.4%
Gold vs. annotator	0.13	0.20	0.17	25.2%	31.9%	28.2%

Table 2: Statistics of the pairwise inter-annotator agreement

Gold	#annotators who assigned the gold label						
emotion	0	1	2	3	4	5	6
ANGER	31.4%	14.3%	5.7%	14.3%	5.7%	17.1%	11.4%
DISGUST	37.1%	20.0%	5.7%	8.6%	5.7%	8.6%	14.3%
FEAR	71.4%	17.1%	5.7%	2.9%	2.9%	0.0%	0.0%
JOY	42.9%	17.1%	8.6%	17.1%	5.7%	0.0%	8.6%
SADNESS	25.7%	25.7%	11.4%	5.7%	11.4%	11.4%	8.6%
SURPRISE	45.7%	22.9%	5.7%	14.3%	5.7%	2.9%	2.9%
all	42.4%	19.5%	7.1%	10.5%	6.2%	6.7%	7.6%

Table 3: The percentage of instances on which certain number of annotators assigned the gold label.

gold labels, are presented in Table 2. As can be seen, the minimum, maximum, and average pairwise inter-annotator agreement is notably lower between any annotator and the gold labels than between any two annotators. The Cohen's κ score between any annotator and the gold label corresponds to only a *slight* (strength of) agreement ($0.00 \leq \kappa \leq 0.20$), while the Cohen's κ for any pair of annotators range from *fair* ($0.21 \leq \kappa \leq 0.40$) to *moderate* ($0.41 \leq \kappa \leq 0.60$) agreement, according to Landis and Koch (1977).

To investigate whether the annotation disagreements stem from poor quality of annotations, or from the natural complexity of the task, all labels assigned by all six annotators were manually checked. It was found that none of the annotators had more than 1% of erroneous annotations. The found errors were either due to labelling topic of the post instead of the writer's emotion (SURPRISE labels), or due to labelling posts based on the words that occur in them instead of the writer's emotion (SADNESS labels). The rest of the annotation disagreements were the result of the natural complexity of the task (Section 4.1.2).

The fairly low agreement among the annotators indicates that the task of detecting Ekman's emotions in tweets is challenging and/or subjective. This is in line with previous studies which reported that emotion detection from text is a complex task that results in low inter-annotator agreements regardless of the emotion taxonomy used (Alm et al., 2005; Schuff et al., 2017; Kim and Klinger, 2018; Bostan and Klinger, 2018; Öhman et al., 2020; Acheampong et al., 2020).

The very low agreement between the annotators and the gold labels, in turn, indicates potential problems with the strategy of automatically assigning gold labels for emotion in tweets (according to the last hashtag of the tweet). This is in line with the results reported by Demszky et al. (2020) where the transfer-learning based system obtained noticeably lower results on the TEC dataset (F_1-score of $\approx$0.5) than on the other two Twitter datasets where the gold label were obtained by manual annotations (F_1-scores of $\approx$0.8 and $\approx$0.7).

4.1.1 Reliability of the Gold Labels

To explore the main causes of disagreements of the annotators with the automatically assigned gold labels, the percentage of instances on which certain number of annotators assigned the same label as the gold one was calculated for each emotion category separately (Table 3). In as many as 42.4% of the cases, none of the six trained annotators assigned the same label as the gold one. All six annotators assigned the same label as the gold one in only 7.6% of the cases. While the latter can be a consequence of a high subjectivity of the task, the former indicates that the automatically assigned gold labels might not be reliable.

In the per-class analysis (Table 3), FEAR and SURPRISE stand out as gold labels for which in as many as 71.4% and 45.7% of the cases, respectively, none of the six annotators assigned the gold

#	Example	Gold	Assigned
1	Relatives here. Hafta sleep on a couch in the basement. #cantsleep #effuguysiwantmyqueensize	FEAR	ANGER(3), SADNESS(20, NEUTRAL(1)
2	There is dirty underwear on the floor of the Men's room in Dillons.	FEAR	DISGUST(6)
3	Sometimes in life u just have to DO IT. holds people back from doing so many things!	FEAR	ANGER(2), SADNESS(2), NEUTRAL(2)
4	Courage is the path that leads from to action. Christian McCormack #quote #quotes	FEAR	NEUTRAL(5), JOY(1)
5	My team is starting to heat up you can't contain us too long let the blowout begin ducks attack the duck	FEAR	ANGER(3), JOY(2), NEUTRAL(1)
6	Wanna be remembered? On black friday, go to a store, push a kid over, look him in the eye and say "You remember me"	FEAR	ANGER(2), NEUTRAL(2), SADNESS(1), JOY(1)
7	I like doing stuff for my close friends when they don't expect it for @lexi_peters	SURPRISE	JOY(6)
8	Looking forward to get this done and seeing the reaction from my beautiful gf if you ask I won't say what it is #happy"	SURPRISE	JOY(6)

Table 4: Examples of complete disagreement of annotators with the gold labels FEAR and SURPRISE. The number in parenthesis after the label signifies the number of annotators who assigned that label.

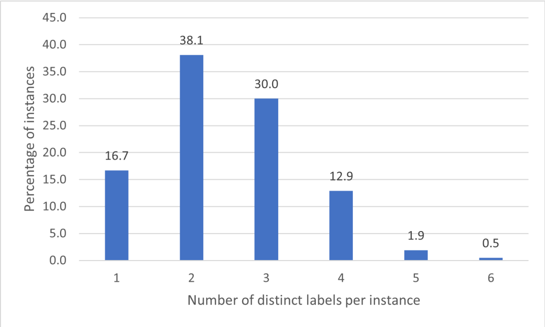

Figure 1: The frequency of obtaining n distinct labels.

label, and in 0.0% and 2.9% of the cases only, all six annotators assigned the gold label. To better understand those phenomena, all instances with FEAR and SURPRISE gold labels, and their annotations by the six annotators, were manually checked.

Among 26 examples with the gold label FEAR, for which none of the six annotators assigned that label, in only five examples (19.2%) it was possible to find some indications of the writer experiencing *fear*, though the emotion labels assigned by the six annotators seemed more probable (examples 1 and 2, Table 4). Another five examples were found where FEAR seems more likely to be the topic than the emotion that the writer experiences, or the post evokes (examples 3 and 4, Table 4). In another five of those 26 cases, FEAR seems to denote the emotion that is expected to be evoked in particular group of people who read the post (examples 5 and 6, Table 4). In eight out of 16 examples with the gold label SURPRISE for which none of the six annotators assigned that label, SURPRISE was rather the topic of the post than the writer's emotion (examples 7 and 8, Table 4).

Set of labels	% of instances
{SADNESS, NEUTRAL}	5.7
{JOY, NEUTRAL}	5.2
{JOY}	5.2
{ANGER, SADNESS}	4.8
{JOY, SADNESS, NEUTRAL}	4.8
{ANGER, DISGUST}	4.3
{JOY, SURPRISE}	3.8
{ANGER}	3.3
{ANGER, NEUTRAL}	2.9
{DISGUST}	2.9
{JOY, SURPRISE, NEUTRAL}	2.9
{ANGER, SADNESS, NEUTRAL}	2.4
{ANGER, DISGUST, SURPRISE}	2.4
{ANGER, JOY, NEUTRAL}	2.4
{ANGER, DISGUST, SADNESS}	2.4
{JOY, SADNESS}	2.4
{SADNESS, SURPRISE}	2.4
{ANGER, SURPRISE}	2.4

Table 5: Most frequently assigned sets of labels.

4.1.2 Inter-Annotator Disagreements

To better understand complexity of the task for human annotators, the percentages of the instances for which six annotators assigned 1, 2, 3, ... 6 distinct labels in total are presented in Figure 1, and the most frequently assigned sets of labels per instance are presented in Table 5. As can be seen, a tweet was most commonly assigned two distinct labels from six annotators. In 37.5% of those cases, the second label was NEUTRAL. The most common

#	Example	Assigned
1	Another evening, another cup of coffee.	NEUTRAL(3), SADNESS(2), JOY(1)
2	At the dentist bright and early	JOY(3), NEUTRAL(2), SADNESS(1)
3	No school, getting up at 8 for a seven hour car ride at least i have #noschool	SADNESS(3), JOY(3)
4	Finally done with work and have to be back in less than 12 hours	SADNESS(5), JOY(1)
5	The movie click is old but one of my favs the ending when he dies makes me tear up	SADNESS(5), JOY(1)

Table 6: Examples that were assigned both JOY and SADNESS. The number in parenthesis after the label signifies the number of annotators who assigned that label.

other disagreements in that group were between ANGER and SADNESS, ANGER and DISGUST, and JOY and SURPRISE (Table 5).

The most surprising combinations, among the most frequently encountered ones, were {JOY, SADNESS} (in 2.4% of all instances) and {JOY, SADNESS, NEUTRAL} (in 4.8% of all instances). By manual inspection of all those surprising combinations, it was discovered that they were not a result of erroneous annotations, but were rather assigned to one of the two following types of posts: (1) posts that can be associated with either *joy* or *sadness* depending on the writer's association with the mentioned action (examples 1 and 2 in Table 6); (2) posts that contain one part that conveys writer's *sadness* and other that conveys writer's *joy* (examples 3–5 in Table 6). Disagreements that stem from the first type of posts cannot be avoided. Disagreements that stem from the second type of posts, in contrast, could be avoided by more fine-grained emotion annotation where annotators also mark the causes for each found emotion in the sentence, as it was done in studies by Kim and Klinger (2018) or Bostan et al. (2020), mentioned in Section 2.

4.2 Comparison of Different Strategies

To systematically analyse the influence of different strategies for obtaining gold label on the manual classification performances, the Cohen's κ and accuracy of each remaining annotator (whose annotations were not used for obtaining the gold labels) against the gold labels were calculated. As the focus of this study is on single-label classification task, the majority vote was used as the gold label (other strategies mentioned in Table 1 would lead to multiple gold labels per instance). As mentioned in Section 3.3, four strategies were explored: automatically obtaining gold labels based on the last hashtag (0 annotations), having one human annotator to provide the gold label (1 annotation), having three human annotators to provide the gold label as the majority label (3 annotations), and having five human annotators to provide the gold label as the majority label (5 annotations). In the last two strategies, it is not always possible to obtain the majority label, i.e. if all three annotators assign different labels (for the strategy where gold labels are obtained based on three human annotations), or if two distinct labels were assigned each by two annotators (for the strategy where gold labels are obtained based on five human annotations). The minimal, maximal, and average percentage of instances (out of 210) that did not have a majority class are given in Table 7.

Annotations for gold	Cases without majority min	max	avg	Data points
3	11.4%	22.9%	17.0%	20
5	11.4%	16.2%	13.5%	6

Table 7: Percentage of instances without majority class.

Cohen's κ and accuracy were computed for each annotator against the gold labels only for those annotators whose annotations were not used for obtaining the gold labels, e.g. if the gold labels were obtained by using annotations of the annotators 1, 2, and 5, then the Cohen's κ and accuracy against the gold labels were calculated only for the annotators 3, 4, and 6. That resulted in 20 data points for the strategy of obtaining gold labels from three human annotations (all combinations of three annotators from a total of six annotators), and six data points for the strategies of obtaining gold labels from one or five human annotations. The cases without majority class (when obtaining gold label from three and five annotations) were excluded.

The influence of the strategy for obtaining gold labels (involving zero, one, three, or five annotators) on the observed manual classification performances (Cohen's κ and accuracy) of the rest of the annotators is clearly visible in Figure 2. As the quality of annotations was previously manu-

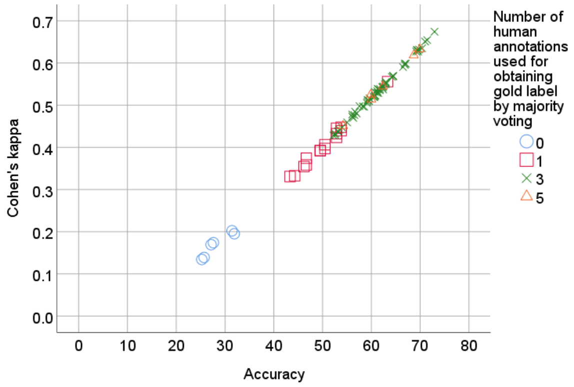

Figure 2: Influence of the strategy for computing gold label on the classification performance.

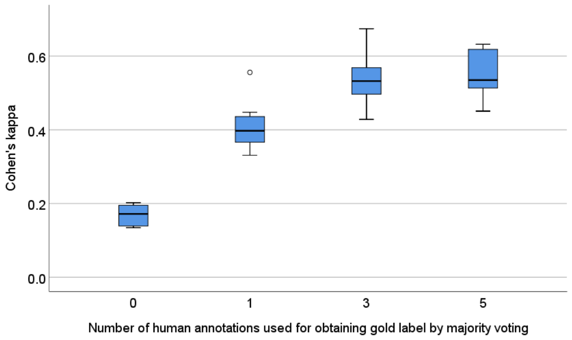

Figure 3: Influence of the strategy for computing gold label on the Cohen's kappa score between the gold label and the rest of the annotators.

ally checked (Section 4.1), the higher classification performances in Figure 2 (the upper right corner of the plot) correspond to better quality of gold labels. According to these results, assigning gold labels based on the majority vote of three or five annotators lead to noticeably higher quality than assigning them based on the annotations of one annotator only. Assigning gold labels automatically, based on the emotion explicitly mentioned in the last hashtag in the tweet, leads to significantly lower quality than any other strategy. Box-plots in Figures 3 and 4 show more detailed results of this analysis. Strategies for obtaining gold labels from three or five human annotations both result in *moderate* to *high* inter-annotator agreements according to Cohen's κ score, and similar average values of those agreements.

The results of this analysis indicate that for obtaining good quality gold labels for Ekman's emotions on English Twitter data it is not necessary to hire more than three trained human annotators. Would the same hold for crowdsourced annotations and different emotion taxonomies is something that needs to be explored in future studies.

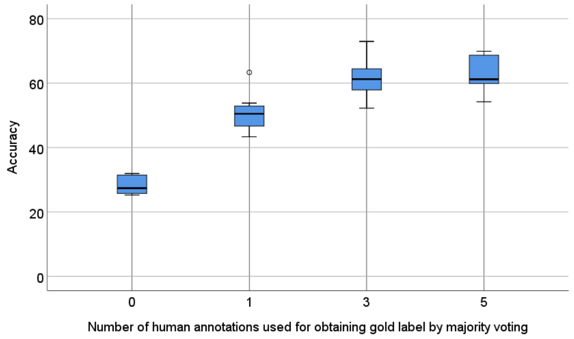

Figure 4: Influence of the strategy for computing gold label on the accuracy of the rest of the annotators.

5 Final Discussion and Conclusion

This study addressed the issue of reliability of gold labels for emotion detection in English tweets.

The results indicated that automatically obtained gold labels (based on the last emotion-hashtag of the tweets) are not reliable, mainly due to the last emotion-hashtag often representing either the topic of the post and not the emotion experienced by the writer, or the emotion that the post is expected to evoke in a particular group of readers. These results call for caution if such large automatically annotated datasets are used for training automatic emotion detection systems, or for testing them, as a significant portion of instances contain suboptimal gold labels.

The analysis of most common disagreements among the annotators revealed that, surprisingly, *joy* and *sadness* are often assigned to the same post by different annotators. A manual inspection of those cases revealed that they are results of either lack of context and knowledge about writer's position about a certain topic, or writer's expression of both sadness and joy in different parts of the post.

The analysis of impact of strategy used for obtaining gold labels on the manual classification results and quality of the test dataset indicated that three trained annotators are sufficient for providing gold labels by their majority vote.

Ethics/Impact Statement

This study is expected to have a broader impact on the field of automatic emotion detection in texts by raising awareness about the complexity of the task, and encouraging other NLP researchers to further explore annotation procedures and the quality of the gold labels in datasets used in automatic emotion detection.

The use of suboptimal annotation schemes and procedures, that do not account for natural complexities of the task, may lead to a high number of incorrect or incomplete labels in compiled datasets. The use of such datasets to train and test NLP models further leads to rewarding the models which are not actually performing well on the final goal but are, instead, good at learning and propagating the errors found in the training datasets (Northcutt et al., 2021). This might be particularly dangerous in the case of automatic emotion detection, as such models might be used in the real-world scenarios for a direct communication with real users, e.g. in empathetic chatbots. If those systems fail to grasp the actual emotional state of the user, especially in the case of individuals who are at at-risk conditions in terms of mental health, they may cause further harms for such users. Furthermore, apart from the accurate recognition of the emotional experiences, such systems need to adequately respond to those users that go through emotional upheavals. Thus, special attention should be paid in the development of proper empathetic reactions of chatbots to prevent the potential harm to vulnerable populations.

Acknowledgements

I would like to express my gratitude to Justyna Hedler for her help in conducting human annotation experiment, Seren Yenikent for valuable discussions about emotion detection in psychology, Francisco Rangel for supporting this work, and the six annotators for the high quality of annotations.

References

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62:2937—-2987.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *Text, Speech and Dialogue*, pages 196–205, Berlin, Heidelberg. Springer Berlin Heidelberg.

Nick Anstead and Ben O'Loughlin. 2014. Social Media Analysis and Public Opinion: The 2010 Uk General Election. *Journal of Computer-Mediated Communication*, 20(2):204–220.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.

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. Association for Computational Linguistics.

Joel Brynielsson, Fredrik Johansson, Carl Jonsson, and Anders Westling. 2014. Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises. *Secur. Informatics*, 3(1):7.

Sven Buechel and Udo Hahn. 2017a. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2017b. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 1–12, Valencia, Spain. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

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

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International*

Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Phil Katz, Matt Singleton, and Richard Wicentowski. 2007. SWAT-MP:the SemEval-2007 systems for task 5 and task 14. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 308–313, Prague, Czech Republic. Association for Computational Linguistics.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

J. Richard Landis and Garry G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33:159–74.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019a. Moel: Mixture of empathetic listeners. *CoRR*, abs/1908.07687.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Zihan Liu, and Pascale Fung. 2019b. Caire: An end-to-end empathetic chatbot. *CoRR*, abs/1907.12108.

Maria L. Loureiro and Maria Alló. 2020. Sensing climate change and energy issues: Sentiment and emotion analysis with social media in the u.k. and spain. *Energy Policy*, 143:111490.

Yukun Ma, Khanh Linh Nguyen, Frank Z. Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

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 and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29.

Saif M. Mohammad, Xiao-Dan Zhu, Svetlana Kiritchenko, and Joel D. Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing and Management*, 51:480–499.

Alena Neviarouskaya, H. Prendinger, and M. Ishizuka. 2009. Compositionality principle in recognition of fine-grained emotions from text. In *ICWSM*.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2010. @AM: Textual attitude analysis model. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 80–88, Los Angeles, CA. Association for Computational Linguistics.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A multilingual dataset for sentiment analysis and emotion detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in virtual agents and robots: A survey. *ACM Trans. Interact. Intell. Syst.*, 7(3).

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark. Association for Computational Linguistics.

Ameneh Gholipour Shahraki and Osmar R. Zaïane. 2017. Lexical and learning-based emotion mining from text.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Happybot: Generating empathetic dialogue responses by improving user experience lookahead. *CoRR*, abs/1906.08487.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

W. Wang, L. Chen, K. Thirunarayan, and A. Sheth. 2012. Harnessing twitter ”big data” for automatic emotion identification. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592.

Peng Wu, Xiaotong Li, Si Shen, and Daqing He. 2020. Social media opinion summarization using emotion cognition and convolutional neural networks. *International Journal of Information Management*, 51:101978.