Identifying Worry in Twitter: Beyond Emotion Analysis

Reyha Verma Christian von der Weth Jithin Vachery Mohan Kankanhalli
School of Computing
National University of Singapore, Singapore
{reyha,chris,jithin,mohan}@comp.nus.edu.sg

Abstract

Identifying the worries of individuals and societies plays a crucial role in providing social support and enhancing policy decision-making. Due to the popularity of social media platforms such as Twitter, users share worries about personal issues (e.g., health, finances, relationships) and broader issues (e.g., changes in society, environmental concerns, terrorism) freely. In this paper, we explore and evaluate a wide range of machine learning models to predict worry on Twitter. While this task has been closely associated with emotion prediction, we argue and show that identifying worry needs to be addressed as a separate task given the unique challenges associated with it. We conduct a user study to provide evidence that social media posts express two basic kinds of worry – normative and pathological – as stated in psychology literature. In addition, we show that existing emotion detection techniques underperform, especially while capturing normative worry. Finally, we discuss the current limitations of our approach and propose future applications of the worry identification system.

1 Introduction

Knowing what an individual or society at large, worry about – e.g., unemployment, health issues, ageing, the rise of AI – is an indicator of people's well-being. Capturing information like the possible source and nature (e.g., type, intensity) of people's worry is used by many governmental agencies[1,2,3] to guide their policy decisions. This can range from minor decisions such as initiation of information campaigns (e.g., to counter false information being spread during the COVID-19 pandemic) to major policy changes like introduction and amendment of laws (e.g., increase of minimum wage, mandatory health insurance, data privacy and fake news). Many private companies (e.g., Toyota) listen to the worries of their customers using techniques like "Voice-of-Customer" (Griffin and Hauser, 1993) as part of their Quality Function Deployment (QFD) process (Toma and Naruo, 2017).

Measuring worry, however, is quite challenging. Traditional approaches rely on time-consuming and costly user surveys and polls (e.g., on a large scale, the World Happiness Report (Helliwell et al., 2020), the Global Risk Report (WEF, 2020)). These surveys, although very well-structured, suffer from some severe limitations. Firstly, being resource-consuming, (large) surveys are generally conducted periodically (e.g., 1-2 times a year). This, in turn, creates knowledge gaps as it is hardly possible to track short-term trends following significant events (e.g., pandemic outbreak, natural disasters, terrorist attacks). Secondly, most of the surveys tend to have a narrow scope with a specific, pre-defined purpose. For example, surveys conducted by the public housing agency are most likely to be limited to the worries of the residents of those societies. Lastly, surveys involving sensitive subjects (e.g., racism, immigration, LGBT rights, abortion, religion, politics) tend to suffer from non-response bias of the participants who might divulge their true opinions – even if they are ensured anonymity – in order to adhere to political correctness.

In contrast, social media provides a platform for individuals to freely and continuously express their thoughts, feelings and experiences as well as to share information with other members of the society. Content on platforms like Twitter is generally public and can be easily collected on a large scale, thereby, making social media mining a promising approach to observe and analyze people's worries. Social media, however, comes with its own set

[1] https://www.reach.gov.sg/participate/public-consultation
[2] https://innovate.mygov.in/dpi-public-consultation/
[3] https://www.hpb.gov.sg/community/national-population-health-survey

Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science, pages 72–82

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

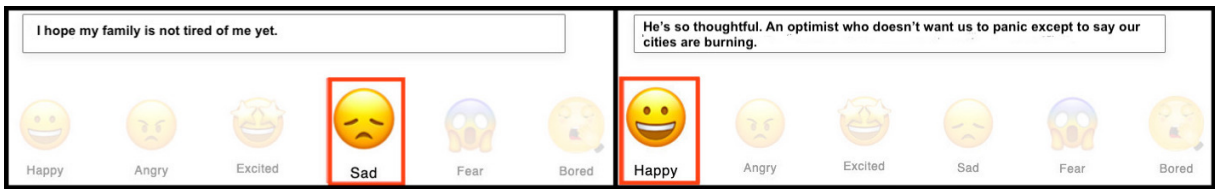

Figure 1: Emotion prediction on tweets with worry.

of challenges in form of self-censorship bias and the presence of scripted bots. Besides this, unlike well-structured surveys that often have structured questions (single or multiple-choice, rating and ranking), social media content is quite unstructured. Sophisticated analytics, most of them using state-of-the-art machine learning algorithms, are often required to extract meaningful and in-depth insights from such unstructured data.

Many works have been proposed to capture people's well-being using social media; see Section 2. In most of these works, worry is associated with notions such as anxiety, fear and nervousness. However, as per psychology literature, worry is primarily considered to be a thought process concerned with future events that often has adverse and uncertain outcomes (Borkovec et al., 1983). Since people are confronted with or affected by events with uncertain outcomes in their day-to-day life, most of them are worried to some extent. Only when the extent of these worries becomes uncontrollable and excessive, this normative process becomes pathological and results in anxiety or depression (Brown, 1997; Watkins, 2008). With the increase in people's worries, their emotional responses tend to get stronger, typically expressed as fear. Thus, while fear implies worry, not all worries necessarily result in fear (Levy and Guttman, 1985). Therefore, we argue that the existing approaches used to evaluate people's mental or emotional states (fear, anxiety, depression, etc.) are not sufficient to accurately capture the notion of worry due to these arguably subtle but important differences. Figure 1 (using ParallelDots[4] API) shows evidence that an emotion classifier is unable to predict worry in tweets classified as sad and happy.

To validate our hypothesis, we first compare people's perception of worry and emotions using the same underlying data. We use crowdsourcing to re-annotate a well-established Twitter dataset curated for emotion prediction for our new target task of identifying worry. We analyze this dataset to establish that emotion is not an adequate predictor of worry. Secondly, using the re-annotated dataset, we train different machine learning models – feature-based, word embedding-based and contextual embedding-based – to refine further the subtleties that result in differences between emotion and worry. Next, we perform an in-depth analysis of different kinds of worry – normative and pathological – by conducting a user study. Lastly, we perform error analysis to highlight current shortcomings as well as challenges while discussing the future work leading to more effective worry prediction.

2 Related Work

Psychology literature defines worry to be a future-oriented thought process typically concerned with a problem whose potentially negative outcome is uncertain (Borkovec et al., 1983). Most people deal with some degree of worry on a daily basis in the form of normative or non-pathological worry (Eysenck, 1995). However, excessive, pervasive and uncontrollable worry becomes pathological in the form of generalized anxiety disorder (Brown, 1997). Besides this difference in intensity of worry, psychology also categorizes worry into various life domains (e.g., health, social relations, environment) as well as the object of worry (e.g., self, close friends/relatives, society and the world) (Boehnke et al., 1998; Schwartz et al., 2000; Schwartz and Melech, 2000).

As social media has become one of the most popular services online, users on these platforms indulge in widespread sharing of thoughts, opinions and feelings, as well as, events that constitute their everyday lives like check-ins, relationships, and more (Schwartz et al., 2013). Several studies have shown that the language, linguistic style and behavior derived from social media posts often reflect users' personal characteristics (Kosinski et al., 2013; Schwartz et al., 2016). Consequently, many methods have been proposed to use social media content to predict users' personality traits

[4]https://www.paralleldots.com/emotion-analysis

and well-being. Most works such as (Azucar et al., 2018; Farnadi et al., 2016; Skowron et al., 2016; Hughes et al., 2012), aim to predict the personality of social media users using the Big 5 personality traits: OCEAN (openness, conscientiousness, extraversion, agreeableness, and neuroticism). While personality traits help in understanding how often a user might worry, they do not allow us to predict whether a particular post expresses worry.

Existing efforts towards evaluating users' well-being related to worry have focused on emotion prediction (Canales, Lea and Martínez-Barco, Patricio, 2014), depression (Guntuku et al., 2017; Choudhury et al., 2013), anxiety and stress (Coppersmith et al., 2014), suicidal thoughts (De Choudhury et al., 2016) – that is, on pathological causes affecting users' well-being. However, worrying is not necessarily pathological and only becomes so when it grows excessive and uncontrollable (Brown, 1997). Similarly, since "daily worries" are part of most people's lives, not all worries trigger a (strong) emotional response. Worry, as a thought process concerned with future events that have uncertain and potentially (very) negative outcomes, is most closely related to fear and anxiety. In fact, many existing works predicting emotions in users' social media posts (Lamb et al., 2013; Harb and Becker, 2018; Wang et al., 2012) associate fear with worry. Since worry does not necessarily imply fear, emotions alone are not a good predictor for worry, as they are skewed towards strong feelings of worry that more likely yield an emotional response.

3 Datasets and Experimental Setup

In this section, we firstly describe our dataset for the task of identifying worry in tweets, then briefly discuss the methodology of data pre-processing and finally outline the set of machine learning models used for evaluation.

3.1 Worry Datasets

For identifying worry in tweets, we leverage on the existing dataset made available to by the SemEval-2018 Task 1: Affect in Tweets (Mohammad et al., 2018), containing 12,634 tweets. This dataset contains four subsets – one for each emotion (joy, fear, anger and sadness). Each of these tweets contains an integer intensity score ranging from 0 to 3 representing no, low, moderate and high intensity respectively. However, there are 1,544 overlapping tweets that are present in more than one subset (e.g., a tweet in "fear" subset with intensity 3 is also present in "joy" subset with intensity 0). We remove these overlapping tweets for the ease of annotation.

Before finalizing the annotation procedure, we conducted three pilot studies with 1,000 tweets per study. We used Amazon Mechanical Turk[5] as the crowdsourcing platform for all the studies. There were 50 annotation tasks each comprising of 20 tweets that were labeled by 5 different native English annotators. Each annotation task included a detailed set of instructions containing examples of different kinds of worries (such as explicit and implicit).

The first pilot study had a 5-point Likert scale with classes: "definitely yes", "probably yes", "not sure", "definitely not" and "probably not" where 'yes' and 'no' referred to the presence and absence of worry respectively. We found that 79.4% of the 1,000 total tweets had no consensus among the workers and, therefore, had to be rejected. Since the task of worry identification is highly subjective, classes containing words 'definitely' and 'probably' added a notion of worry intensity, making it difficult for the workers to be confident of their annotations. Therefore, for our second study, we switched over to the 3-point Likert scale with classes: "worry", "non-worry" and "not sure". The rejection rate drastically reduced to 30.15% as workers became more confident of their assessment. However, on a closer inspection, we found that some annotations to be unsatisfactory due to lack of quality control.

We performed the third pilot study with a 3-point Likert scale with classes: "worry", "non-worry" and "not sure" and a quality control mechanism in the form of test tweets. For quality control, 3 out of every 20 tweets were test tweets manually created by the authors that unquestionably expressed worry or no worry as a result of which the rejection rates further fell down to 16.4%. Finally, we used the last pilot study to label the remaining 11,090 tweets. After discarding tweets labeled as "not sure", we got a total of 10,191 tweets which are contained in our worry Twitter (**WT**) dataset. The rejection rate for the final dataset is 15.14% and the inter-rater agreement is 0.258 ('fair agreement' as per Fleiss Kappa (Landis and Koch, 1977)) Details are given in Table 1.

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

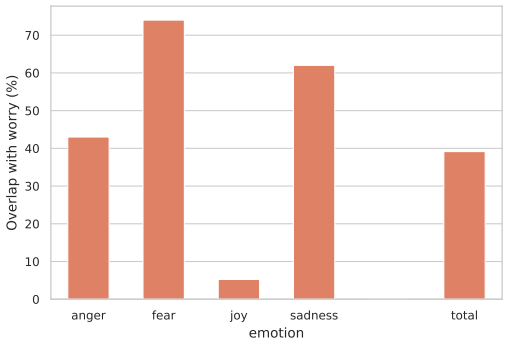

Figure 2: Overlaps of emotion and worry labels.

To compare worry with emotion (cf. 4.2), we create an additional **WT-Short** dataset removing 4,165 tweets with intensity 0 out of the total 10,191 annotated tweets. The original **WT** dataset is used to evaluate our worry classifier whereas the **WT-Short** dataset is used primarily to analyze the differences between worry and emotion.

Dataset	Non-Worry	Worry	Total
WT	6,836	3,355	10,191
WT-Short	3,666	2,360	6,026

Table 1: Details of used datasets.

Having a single **WT-Short** dataset annotated with both worry and emotions allows us to compare the two concepts. Figure 2 shows the overlap between the tweets labeled with worry for each of the four emotions. While, not unexpectedly, the overlap between fear and worry is the highest and the overlap between joy and worry is the lowest, there is no clear connection of other emotions - sadness, anger - with worry.

3.2 Data Pre-processing

We use the Ekphrasis tool (Baziotis et al., 2017) as the pre-processor for our dataset. Ekphrasis recognizes Twitter mark-up, emoticons, emojis, dates, currencies and words with emphasis using an exhaustive list of regular expressions. Using Ekphrasis, we perform Twitter-specific tokenization followed by spell correction and word normalization. For traditional feature-based models, we remove the stop-words and lemmatize the tweets as an additional step.

3.3 Model Description

To carry out the experiments, we explore three different machine learning approaches for worry detection: a traditional approach using feature-based models, deep learning approaches based on non-contextual (word-based) embeddings and deep learning approaches based on contextual embeddings.

For traditional approaches, we use (1) Multinomial Naïve Bayes (MNB), and (2) Support Vector Machine (SVM) implementations available in scikit-learn (Pedregosa et al., 2012). For deep learning approaches using the word-based embeddings, we use (1) Hierarchical Attention Network (HAN) (Yang et al., 2016), which utilizes hierarchical nature of the text along with the attention mechanism, and (2) CNN for sentence classification (CNN-static) (Kim, 2014) which consists of max pooling and convolution. Since emojis are quite frequently used in tweets and convey important information, we combine the GloVe vectors (Pennington et al., 2014) trained on $840B$[6] tokens with 300-dimensional emoji2vec embeddings (Eisner et al., 2016) to ensure that our emojis are also well-represented while training. For deep learning approaches based on contextual embeddings, we use (1) RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et al., 2019) and XL-Net: Generalized Autoregressive Pretraining for Language Understanding (Yang et al., 2019) made available by HuggingFace.[7] We, then, fine-tune these models for our classification task.

For training the models in experiments mentioned in Section 4, we split the dataset into 80-10-10(%) for train-dev-test respectively. We repeat each experiment five times and average the results. For optimizing our network, we use Adam (Kingma and Ba, 2015) with a mini-batch of size 32. We use TensorFlow (Abadi et al., 2016) for implementing all our deep learning models. Details on the hyper-parameter values and the modifications made to the architectures are mentioned in the Appendix.

4 Experiments and Results

This section covers four parts: the evaluation of our worry classifier, the effect of emotion on worry prediction, analysis of different kinds of worry using a user study and an error analysis discussing the core challenges towards further improving this task.

[6]https://nlp.stanford.edu/projects/glove/
[7]https://huggingface.co/transformers/

4.1 Worry Prediction

For this experiment, we train six different models, as discussed in Section 3.3, using both **WT** and **WT-Short** dataset. Table 6 shows the results on the test set using different metrics. Note, we use the Matthew Correlation Coefficient (MCC) as an additional metric because our dataset is highly imbalanced.

Model	Precision	Recall	F1-score	Accuracy	MCC
MNB	0.84	0.08	0.14	0.68	0.19
SVM	0.59	0.46	0.52	0.71	0.32
CNN-static	0.59	0.62	0.60	0.73	0.39
HAN	0.62	0.52	0.57	0.74	0.38
RoBERTa-GRU	0.69	0.55	**0.61**	0.76	0.45
XLNet-GRU	0.73	0.50	0.59	0.75	0.44
Model	**Precision**	**Recall**	**F1-score**	**Accuracy**	**MCC**
MNB	0.79	0.31	0.45	0.68	0.34
SVM	0.67	0.63	0.65	0.72	0.42
CNN-static	0.66	0.78	0.72	0.74	0.41
HAN	0.68	0.74	0.71	0.75	0.43
RoBERTa-GRU	0.66	0.86	**0.75**	0.76	0.5
XLNet-GRU	0.68	0.79	0.73	0.74	0.45

Table 2: Performance of worry classifiers on the **WT** dataset (top) and the **WT-Short** dataset (bottom)

We see that for **WT** dataset, the deep learning models achieve a much higher F1-score as compared to the traditional models. Among different deep learning models, RoBERTa that uses a byte-level BPE (Byte-Pair Encoding) token on top of the BERT (Bidirectional Encoder Representations from Transformers) model outperforms others (F1-score: 61%). Though the score is only slightly higher than the other models, given the nature of the task, these results are not surprising as worry identification requires in-depth contextual knowledge. On careful analysis of the false positives and false negatives, we observe that these tweets are more ambiguous and subjective, making them more difficult to classify. We detail these challenging cases in Section 4.4

The six additional classifiers that are trained using the **WT-Short** dataset to compare worry and emotion as specified in Section 4.2. This dataset contains tweets with emotion intensity equal to or greater than one making it possible for us to perform comparative analysis. It is because the **WT-Short** dataset is more balanced as compared to the **WT** dataset that we find a clear difference between the results obtained, as shown in Table 6.

4.2 Worry with/vs. Emotion Prediction

To analyze the relationship between worry and emotion, we perform two additional series of experiments. Firstly, we evaluate if emotion and sentiment improve the task of identifying worry. As worry is often assumed to be related to negative future events (Borkovec et al., 1983), we additionally use sentiment labels. The main objective of analyzing sentiment is to understand if polarity, particularly negative polarity, plays any role in aiding worry detection task.

In order to do this, we train three classifiers – SVM, CNN-static and RoBERTa-GRU – the best performing models in their respective categories (cf. Table 6). For each of these three classifiers, we train four combinations with the following inputs: (1) worry (W) (2) worry and sentiment (W+S) (3) worry and emotion (W+E) (4) worry, emotion and sentiment (W+E+S) to be able to perform an in-depth analysis. To obtain the sentiment labels, we use VADER (Hutto and Gilbert, 2014), a sentiment analyzer optimized for social media content such as tweets. For deep learning models, sentiment and emotion annotations are added as input layers before the dense output layer. For traditional models, input features are simply concatenated together.

Model	Input	Precision	Recall	F1-score	Accuracy	MCC
SVM	W	0.67	0.63	0.65	0.72	0.42
	W+S	0.69	0.66	0.68	0.71	0.41
	W+E	0.70	0.67	0.69	0.75	0.48
	W+E+S	0.69	0.67	0.68	0.74	0.47
CNN-static	W	0.66	0.78	0.72	0.74	0.41
	W+S	0.64	0.73	0.68	0.72	0.46
	W+E.	0.67	0.84	0.74	0.75	0.47
	W+E+S	0.68	0.75	0.72	0.74	0.52
RoBERTa-GRU	W	0.66	0.86	0.75	0.76	0.50
	W+S	0.66	0.76	0.71	0.74	0.50
	W+E	0.72	0.78	0.75	0.79	0.50
	W+E+S	0.71	0.75	0.73	0.77	0.47

Table 3: Performance of the worry classifiers and joint classifiers on the **WT-Short** dataset. Here, W = worry, W+S = worry and sentiment, W+E = worry and emotion, W+E+S = worry, emotion and sentiment.

Table 3 shows the results over the **WT-Short** dataset. We observe that for all the four different combinations, RoBERTa-GRU performs better than the rest. Although emotion does not improve worry prediction to a great extent, there is a slight increase in the scores obtained. For RoBERTa-GRU, the accuracy after adding the emotion inputs (W+E) improves from 76% to 79%. Note, this does not indicate emotion is an equally good predictor of worry. The F1-score for both W+S and W+E+S decreases, thereby suggesting a negative impact of sentiment on the results.

Secondly, we evaluate how well an emotion classifier can serve as a predictor of worry. For example, if a tweet is classified as “fear” or “sadness”, how good is the prediction with respect to worry.

For this experiment, we first perform a logistic regression analysis using emotion labels as inputs to predict worry. The coefficients for fear, anger, sadness and joy are 0.445, 0.035, 0.429 and -1.55 respectively. As expected, fear and joy have the highest and the lowest coefficients respectively. Next, we move on to text modeling and train an emotion classifier using **WT-Short** dataset as shown in Table 4. We use the best model for emotion classification (RoBERTa-GRU) to evaluate the results for all possible combinations of the four emotions. Figure 3 shows the corresponding F1-scores for the worry classification task ranked from best to worst. Each combination represents one class, and each class consists of one or more emotions combined together to predict worry.

Model	Fear	Anger	Sadness	Joy
SVM	0.73	0.77	0.68	0.71
CNN-static	0.78	0.80	0.73	0.92
RoBERTa-GRU	0.80	0.85	0.76	0.94

Table 4: F1-scores of emotion classifiers trained on **WT-Short** dataset.

As intuitively expected, the best emotion combinations contain "fear", although "fear" on its own performs quite low. Given the relationship between worry and emotions, it is not surprising that emotion classification can serve as a predictor for worry, although subpar to our worry classifier. However, we also argue that the results in Figure 3 represent a best-case scenario since we built upon a dataset created for emotion classification. In the following sections, we show that worry is often only implied in a neutral, "emotionless" manner, making an emotion classifier generally unsuitable for worry prediction.

4.3 User Study

We evaluate the ability of our classifiers to clearly distinguish between pathological and normative worry by conducting an empirical study using a non-emotion dataset. The reason for selecting a non-emotion dataset is to obtain sufficient tweets containing normative worry as emotion datasets are usually dominated by tweets with pathological worry. To compare worry with emotion, we consider the combination of three negative emotions – i.e. "fear", "sadness" and "anger" – as worry and the positive emotion "joy" as non-worry throughout this section. It is because these combinations are found to be the best and the worst predictors of

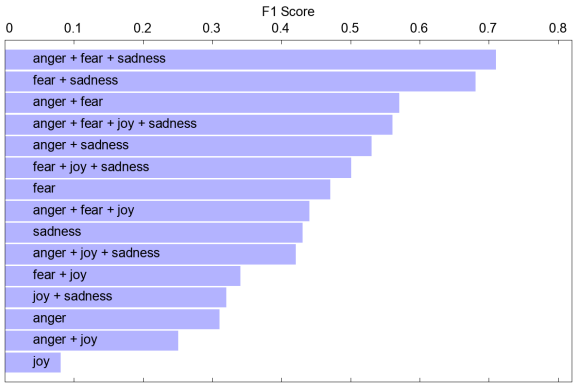

Figure 3: F1 scores of different emotion classifiers for the prediction of worry.

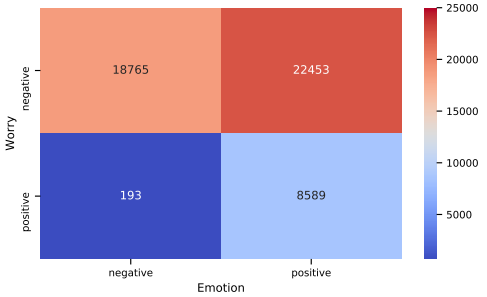

Figure 4: Worry vs. Emotion using **Sentiment140** dataset.

worry respectively (cf. Figure 3). In this section, we will first describe the setup for our study and then discuss the results obtained.

4.3.1 Experimental Setup

For setting up this experiment, we use the **Sentiment140** dataset provided by Go et al. (2009). This dataset is a corpus for sentiment analysis that contains 1.6 million tweets annotated with polarity 0 (negative) and 1 (positive). Out of the 1.6 million tweets, we randomly sample 50k tweets containing an equal number of tweets with positive and negative sentiment. Using the classifiers discussed in Section 4, we predict emotion and worry for these 50k tweets. Figure 4 shows the overall distribution. We then annotate 500 tweets - 193 worry tweets with negative emotion and 307 worry tweets with positive emotion (randomly sampled out of approximately 8k tweets). We follow the same annotation process, as the one discussed in Section 3.1 and obtain worry labels to validate our worry predictions.

4.3.2 Normative vs. Pathological Worry

Next, we examine the annotated dataset for normative and pathological worries. Table 5 shows different normative and pathological tweets. Normative tweets are the worry tweets correctly predicted by our worry classifier but incorrectly predicted by our emotion classifier. The pathological tweets, on the other hand, are the worry tweets correctly predicted by both our worry and emotion classifier.

Pathological tweets like *"im pissed!!as in very very pissed!!"* or *"have my phone interview for DCU today.... soooooooooooooo nervous!!!"* are the tweets where users explicitly mention they are experiencing a negative emotion and/or are often associated with high emotional intensity. These are mostly correctly predicted as worry by both worry and emotion classifier. This is, however, not true for normative worries. While evaluating normative worries, we find there are two different kinds of tweets.

Mixed Emotions. The most frequent scenario is where a tweet expresses mixed emotions, that is, both positive and negative emotions simultaneously. For example, *"Home alone again.....YAY!!!! BUT STILL HAVE TO DO MORE CHORES.....UGH."*. The first half of this sentence conveys happiness, while the second half expresses sadness. The emotion classifier, being unable to capture the underlying worry, predicts this to be joy (non-worry). The emotion classifier usually performs poorly on such tweets. To correctly classify worry tweets with mixed emotions, it is imperative to train the model on dataset that is specifically curated for the task of worry detection.

Low or Zero Emotional Intensity. Social media users use different writing styles to express themselves. For example, *"Wow only got 3 hrs of sleep!! Bad Bad bad!!!! got a huge headache!!!"* represents an extremely intense tweet with high intensity of negative emotion, making it easier for our emotion classifier to classify it as worry. However, tweets having an extremely low or zero emotional intensity such as *"I really wish my life was a little more exciting"* are most likely to not get captured by the emotion classifier. It is because of these subtle differences in the nature of the task of emotion detection and worry detection that we argue the need to study worry detection as a separate problem.

When using emotion classifier as a predictor of worry, the chances of missing out on capturing worry, especially normative worry, in tweets is quite high due to above-mentioned differences. This shows that the task of worry classification can be considerably improved when studied separately.

4.4 Error Analysis and Discussion

The results in Section 4.1 show that identifying worry is a non-trivial task. In this section, we perform a qualitative error analysis by looking at all false positives and false negatives in order to identify the common causes of incorrect classifications.

Expressed vs. Implied Worry. The most prominent case where a tweet is labeled with worry but is not correctly classified is when worry is not explicitly expressed but only implied. For example, consider the tweet *"My parents just had a car accident."* On a purely syntactic level, this tweet does not express a worry. However, knowing that car accidents are generally associated with adverse outcomes such as injury (or death) and financial burden – and also granting the writer empathy – it is very likely that the writer is indeed worried. That worry which is only implied but not (strongly) expressed is particularly common for everyday worries such as issues at home or work that are not severe enough to cause a strong (emotional) response that would more likely reflect in the tweet. This distinction between expressed and implied worry is an instance of a fundamental challenge for NLP: the difference between what a text states and the full message conveyed by the text in the context of shared and common knowledge (e.g., traffic accidents often result in serious injuries). While methods to incorporate external knowledge for text classification have been proposed (Wang et al., 2017; Chen et al., 2019), they work on concept hierarchies (e.g., accident *is-a* misfortune *is-a* event) and it is not apparent how they can sufficiently capture the notion of worry.

Emotions "Hiding" Worry. In contrast to psychological literature that states that any fear implies a worry (Levy and Guttman, 1985), our dataset contains many tweets labeled as "fear" but not as "worry" (similar for "sadness" and "anger"). When inspecting those tweets, we observed that in many cases, the emotion was very obvious like in *"I..I can't! I'm scared! Bees terrify me"* or *"I can never find the exact #emoji that I'm after at the exact moment that I need it #panic"*, both labeled with "fear" but not with "worry". Our explanation is that

Type	Tweets
Normative	I really wish my life was a little more exciting
	Awesome day again, shame the good weather will be gone by the weekend
	sick and in good spirits. its only a sore throat.
	Traffic on a beautiful day where the hell is everyone going
	Sooo happy to be home but it's bittersweet because my wife, son and dog aren't here
	Home alone again.....YAY!!!! BUT STILL HAVE TO DO MORE CHORES.....UGH
	Rain is cool until it starts leaking into your house, ruining stuff.
	I'm also suddenly not feeling well. That's fun.
	Its been a slow day at home, one of my kids is sick. This little picture cheered me up URL
	It's beautiful outside! But I'm stuck inside doing homework
Pathological	im pissed!!as in very very pissed!!
	Goddammit..I'm in trouble
	have my phone interview for DCU today.... sooooooooooooooo nervous!!!!
	Oh my god my head hurts so damn bad! I wanna sleep!
	WOOOOAAAWWW I'M A HORRIBLE STUDENT!!!! AND THE SINGLE WORST PROCRASTINATOR IN THE WHOLE WIDE WORLD!!!!!!!
	Wow only got 3 hrs of sleep!! Bad Bad bad!!!! got a huge headache!!!
	Im kinda nervous for this orientation
	Feeling terrible. Why isn't the day over yet?
	Still not asleep. Ahhh Wtf?!
	MY MOM NEVER CAME HOME AND CALLED REALY EARLY BUT I WAS ASLEEP AND NOW SHE WONT ANSWER THE PHONE AND SHE IS NOT AT WORK I AM SCARED!

Table 5: Normative and Pathological worry tweets obtained using **Sentiment140** dataset.

a strong and explicit expression of emotion, particularly fear, may distract from the underlying cause such as worry. Furthermore, while these tweets have negative sentiment, they do no express or imply an uncertain outcome of a future event, making them less likely to be associated with worry by a reader. This subjectivity is a fundamental issue and may only be addressed adequately in the context of a specific application scenario.

Informal Writing and Stylistic Devices. Lastly, as for most NLP tasks over social media content such as tweets, our worry classifier suffers not only from the informal writing style but also from the often used stylistic devices. Despite careful preprocessing of the tweets, typos, non-standard abbreviations, Internet slang, expressive lengthenings, etc. – e.g., *"I start work tmrw yall, I'm neeervous lol"* – negatively affect the learning and prediction process. Stylistic devices such as sarcasm, irony or humor make it very difficult, even for humans, to assess whether a worry (or emotion or sentiment) is sincere. *"I absolutely love having an anxiety attack halfway through a family meal"* and *"I want my diamonds as bright as my future"* are two examples for this. Existing works towards, e.g., sarcasm detection (Bamman and Smith, 2015; Rajadesingan et al., 2015) or irony detection (Reyes et al., 2013) might help in the long run to further improve the identification of worry in tweets.

5 Conclusion and Future Work

Worry about a personal issue such as health or finance, or a broader external issue such as environmental pollution, technology change or social structure is commonly expressed on Twitter nowa-days. Most of the existing works utilizing social media for measuring well-being associate worry with emotions. Taking cues from the psychology literature that clearly differentiates pathological (uncontrollable with high emotion intensity) and normative (everyday with low/zero emotion intensity) worry, we argue as to why emotion classifier is unable to capture normative worry, thereby, establishing the need to treat worry detection as a separate task.

We started out by exploring the effectiveness of a worry classifier by comparing different state-of-the-art text classification models with/vs. emotion. We then conducted an empirical user study to further strengthen our hypothesis where we discussed the differences between pathological and normative worry in detail. Our results support our argument – that emotion classification can, at best, only be sufficient to predict pathological worries as they yield

strong emotional responses. This topic is, however, less explored despite the immense potential in applications such as identifying day-to-day worries like excessive traffic on a certain route or stressful work environment.

This paper lays down an initial ground for future work in this direction but is far from perfect. We highlight the current limitations of this task by performing a qualitative error analysis. One of the main challenges is that worry is often only implied and requires access to shared or common knowledge. Utilizing such knowledge will be an important next step to improve the identification of worry. Looking at the bigger picture, we envision to implement a real-time, automated worry classification system capable of capturing both pathological and normative worries at different levels – local, national and global – to aid policy and decision making processes of organizations all around the world.

Acknowledgments

The authors gratefully acknowledge the research grant (R-252-000-A47-133) provided by Lloyd's Register Foundation Institute for the Public Understanding of Risk. This research is also supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the funding agencies.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian J. Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Józefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Gordon Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul A. Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda B. Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467.

Danny Azucar, Davide Marengo, and Michele Settanni. 2018. Predicting the Big 5 Personality Traits from Digital Footprints on Social Media: A Meta-Analysis. *Personality and Individual Differences*, 124:150–159.

David Bamman and Noah Smith. 2015. Contextualized Sarcasm Detection on Twitter. In *International AAAI Conference on Web and Social Media*.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada. Association for Computational Linguistics.

Klaus Boehnke, Shalom Schwartz, Claudia Stromberg, and Lilach Sagiv. 1998. The Structure and Dynamics of Worry: Theory, Measurement, and Cross-National Replications. *Journal of Personality*, 66(5):745–782.

T.D. Borkovec, Elwood Robinson, Thomas Pruzinsky, and James A. DePree. 1983. Preliminary Exploration of Worry: Some Characteristics and Processes. *Behaviour Research and Therapy*, 21(1):9–16.

Timothy A Brown. 1997. The Nature of Generalized Anxiety Disorder and Pathological Worry: Current Evidence and Conceptual Models. *The Canadian Journal of Psychiatry*, 42(8):817–825. PMID: 9356769.

Canales, Lea and Martínez-Barco, Patricio. 2014. Emotion detection from text: A survey. In *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, pages 37–43. Association for Computational Linguistics.

Jindong Chen, Yizhou Hu, Jingping Liu, Yanghua Xiao, and Haiyun Jiang. 2019. Deep Short Text Classification with Knowledge Powered Attention. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 6252–6259. AAAI Press.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. In *International AAAI Conference on Web and Social Media (ICWSM)*. The AAAI Press.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60. Association for Computational Linguistics.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 2098–2110, New

York, NY, USA. Association for Computing Machinery.

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. Association for Computational Linguistics.

Michael W. Eysenck. 1995. Worrying: Perspectives on Theory, Assessment and Treatment. Edited by G. C. L. Davey and F. Tallis. (Pp. 311; £24.95.) Wiley: Chichester. 1994. *Psychological Medicine*, 25(2):431–432.

Golnoosh Farnadi, Geetha Sitaraman, Shanu Sushmita, Fabio Celli, Michal Kosinski, David Stillwell, Sergio Davalos, Marie-Francine Moens, and Martine Cock. 2016. Computational Personality Recognition in Social Media. *User Modeling and User-Adapted Interaction*, 26(2–3):109–142.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *Processing*, pages 1–6.

Abbie Griffin and John R. Hauser. 1993. The voice of the customer. *Marketing Science*, 12(1):1–27.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting Depression and Mental Illness on Social Media: An Integrative Review. *Current Opinion in Behavioral Sciences*, 18:43–49. Big data in the behavioural sciences.

Jonathas G. D. Harb and Karin Becker. 2018. Emotion Analysis of Reaction to Terrorism on Twitter. In *Proceedings of the SBC Brazilian Symposium on Databases*, pages 97–108. Association for Computational Linguistics.

John Helliwell, Richard Layard, and Jeffrey Sachs. 2020. World Happiness Report 2020, New York: Sustainable Development Solutions Network.

David John Hughes, Moss Rowe, Mark Batey, and Andrew Lee. 2012. A tale of Two Sites: Twitter vs. Facebook and the Personality Predictors of Social Media Usage. *Computers in Human Behavior*, 28(2):561–569.

Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*. The AAAI Press.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private Traits and Attributes are Predictable from Digital Records of Human Behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

Alex Lamb, Michael J. Paul, and Mark Dredze. 2013. Separating Fact from Fear: Tracking Flu Infections on Twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Shlomit Levy and Louis Guttman. 1985. Worry, Fear, and Concern Differentiated. *Issues in Mental Health Nursing*, 7(1-4):251–264.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2012. Scikit-learn: Machine learning in python. *CoRR*, abs/1201.0490.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 97–106, New York, NY, USA. Association for Computing Machinery.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Lang. Resour. Eval.*, 47(1):239–268.

H. Schwartz, Maarten Sap, Margaret Kern, Johannes Eichstaedt, Adam Kapelner, MEGHA AGRAWAL, EDUARDO BLANCO, LUKASZ DZIURZYNSKI, GREGORY PARK, David Stillwell, MICHAL KOSINSKI, Martin Seligman, and Lyle Ungar. 2016. Predicting individual well-being through the language of social media. pages 516–527.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, 8(9).

Shalom H Schwartz and Gila Melech. 2000. National differences in micro and macro worry: Social, economic, and cultural explanations. *Culture and subjective well-being*, pages 219–256.

Shalom H. Schwartz, Lilach Sagiv, and Klaus Boehnke. 2000. Worries and Values. *Journal of Personality*, 68(2):309–346.

Marcin Skowron, Marko Tkalčič, Bruce Ferwerda, and Markus Schedl. 2016. Fusing Social Media Cues: Personality Prediction from Twitter and Instagram. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 107–108, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Sorin-George Toma and Shinji Naruo. 2017. Total quality management and business excellence: The best practices at toyota motor corporation. *Amfiteatru Economic Journal*, 19(45):566–580.

Jin Wang, Zhongyuan Wang, Dawei Zhang, and Jun Yan. 2017. Combining Knowledge with Deep Convolutional Neural Networks for Short Text Classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 2915–2921. AAAI Press.

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 Conference on Social Computing*, pages 587–592. Association for Computational Linguistics.

E. Watkins. 2008. Constructive and unconstructive repetitive thought. *Psychological Bulletin*, 134:163 – 206.

World Economic Forum WEF. 2020. The Global Risks Report 2020.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489. Association for Computational Linguistics.

A Appendices

A.1 Details of Hyper-parameters

Model	Hyper-parameter	Value
SVM	kernel	linear
	decision function shape	ovo (one-vs-one)
	regularization parameter (C)	2
CNN-static	hidden dimension	100
	number of filters	10
	max sequence length	300
	filter size	(3,8)
	dropout probability	(0.3, 0.5)
HAN	max words per sentence	15
	word encoding dimension	200
	sentence encoding dimension	200
	max sequence length	300
RoBERTa	dropout probability	0.2
	attention dropout probability	0.2
	hidden dimension	64
	max sequence length	300
XLNet	dropout probability	0.2
	attention dropout probability	0.2
	hidden dimension	64
	max sequence length	300

Table 6: Details of hyper-parameters

A.2 Used vs. Original Architecture

1. For CNN-static, 10 filters were used instead of original 100, 2 filter sizes instead of 3, 100 hidden dimensions instead of 50, max pooling instead of global pooling.
2. For RoBERTa and XLNet, pre-trained embeddings followed by two stacked Bidirectional GRU layers and a dense layer.