Does Optical Character Recognition and Caption Generation Improve Emotion Detection in Microblog Posts?

Roman Klinger

Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
roman.klinger@ims.uni-stuttgart.de

Abstract. Emotion recognition in microblogs like Twitter is the task of assigning an emotion to a post from a predefined set of labels. This is often performed based on the Tweet text. In this paper, we investigate wether information from attached images contributes to this classification task. We use off-the-shelf tools to extract a signal from an image. Firstly, with employ optical character recognition (OCR), to make embedded text accessable, and secondly, we use automatic caption generation to generalize over the content of the depiction. Our experiments show that using the caption only slightly improves performance and only for the emotions *fear*, *anger*, *disgust* and *trust*. OCR shows a significant impact for *joy*, *love*, *sadness*, *fear*, and *anger*.

Keywords: emotion classification, social media, caption generation, optical character recognition

1 Introduction

In natural language processing, emotion recognition is the task of associating words, phrases or documents with predefined emotions from psychological models. We consider discrete categories as proposed by Ekman [1] and Plutchik [2], namely *anger*, *fear*, *joy*, *sadness*, *surprise*, *disgust*, *love*, *shame*, and *trust*.

Emotion detection has been applied to, *e.g.*, tales [3], blogs [4], and as a very popular domain, microblogs on Twitter [5]. The latter in particular provides a large source of data in the form of user messages [6]. A common source of weak supervision are hashtags and emoticons to train classifiers. These measure the association of all other words in the message with the emotion [7]. For instance the Tweet "Be prepared for a bunch of depressing tweets guys because I am feeling it tonight. Very sad." is associated via the trigger words "depressing" and "sad" with the emotion *sadness* [8]. In this example, a standard text classification approach is likely to succeed. However, there are (at least) two cases in which the expressed emotion is not communicated in the text of the Tweet: Firstly, the main message can be hidden as text in an image which is linked to the post. This is a popular strategy, given constraints on post lengths (for instance on

The final publication is available at Springer via https://doi.org/10.1007/978-3-319-59569-6_39

Twitter to 140 characters) and to advertise a specific message or content with graphical support. In addition, some authors try to hide their message from search engines and platform curators, for instance in hate speech or fake news. Secondly, an author might show their emotion by posting a photo that depicts a particular situation. In this paper, we investigate if the recognition of emotional content of a micropost can be improved by taking into account information from linked images. To achieve this, we extract textual characterizations of the images associated with posts, using off-the-shelf tools.

Related work in this area aims, for instance, at detecting and understanding of facial expressions [9,10]. Similar techniques have also been applied to detection of specific classes of linked images, for instance fake pictures [11]. Recent research showed sentiment classification benefits from the combination of text and image information [12].

We perform experiments on a Twitter corpus with a subset of emotions. For linked emotions, we incorporate two different approaches to extract information and represent it as text: (1) We process every image linked to a Tweet with optical character recognition (OCR) and combine these textual features from the recognized text with the Tweet text. The hypothesis is that OCR text complements the Tweet text and therefore improves classification performance. (2) We process every image with a pretrained neural network-based caption generation model and combine the generated text with the Tweet. We expect that some content is more likely to be associated with specific emotions, for instance groups of people or "selfies" might be more likely to be joyful. In addition, we analyze the corpus and point out interesting future research directions. Our corpus will be publicly available at http://www.ims.uni-stuttgart.de/data/visualemotions.

2 Methods and Experimental Setup

2.1 Features

We formulate the task of emotion detection from microposts as a multiclass classification problem, *i.e.*, we assume each instance belongs to exactly one emotion; for each, we use one maximum entropy classification model. Features are drawn from three different sources: the micropost text itself, text detected by optical character recognition in a linked image, and a generated caption which describes the content of the image.

Micropost text. We represent the text from the Tweet itself as standard bag-of-words. For better reproducibility of the experiments, we only perform token normalization by disregarding all non-alphanumeric characters and the hash sign ("#"). Hashtags which denote an emotion are ignored (see Section 2.2). Usernames (*i.e.*, character sequences starting with @) are mapped to "@USERNAME". We refer to such features of post d as ϕ^d_{text}.

Optical Character Recognition. Each image which is linked to a Tweet is processed by an optical character recognition (OCR) system. We use Tesseract 3.04.01 [13] with default parameters and ignore all output shorter than six

Table 1: Corpus Statistics. D_{all} is sampled from all Tweets, "w/ ϕ^d_{OCR}" and "w/ ϕ^d_{Vis}" denote counts for subsets with OCR features and caption features. D_{OCR} and D_{Vis} are sampled from the set of all Tweets with recognized text in the image and do not contain recognized text in the image.

Emotion	D_{all}	w/ ϕ^d_{OCR}	w/ ϕ^d_{Vis}	D_{OCR}	D_{Vis}
joy	91,836	6,927 (8 %)	18,672 (20 %)	92,066	111,604
love	41,470	3,477 (8 %)	8,963 (22 %)	46,290	50,974
sadness	26,521	1,495 (6 %)	2,952 (11 %)	19707	14,370
fear	12,721	1,490 (12 %)	2,299 (18 %)	19,400	7,925
anger	11,902	831 (7 %)	1,384 (12 %)	10,379	5,317
shame	4,562	138 (3 %)	193 (4 %)	1,988	862
surprise	5,492	195 (4 %)	792 (14 %)	2,790	5,681
trust	5,170	561 (11 %)	886 (17 %)	7,274	3,151
disgust	326	8 (2 %)	22 (7 %)	106	116
All	200,000	15,122 (8 %)	36,163 (18 %)	200,000	200,000

bytes. Features are generated from the recognized text as bag-of-words. In addition, we add one feature which always holds if any text was recognized. We refer to this feature set as ϕ^d_{OCR}.

Caption Generation. Recently, several approaches have been proposed to generate caption-like descriptions of the content of an image [14–16]. Such methods have been shown to be robust enough to serve as off the shelf tools. In these experiments, we rely on NeuralTalk2 (https://github.com/karpathy/neuraltalk2), a deep neural network CNN and RNN architecture. We use the pretrained COCO data set [17] model which is available with NeuralTalk2. As in the OCR output, we add features using a bag-of-words scheme and one feature which indicates whether a caption was generated (*i.e.*, this feature represents only that an image is attached, not the content). We refer to these features as ϕ^d_{Vis}.

2.2 Corpus

To analyze the impact of each feature set and combination, we downloaded Tweets for a particular set of hashtags from Twitter between March and November 2016. The emotions with example hashtags are *joy* (*e.g.* #happy, #joy, #happiness, #glad), *sadness* (*e.g.* #sad, #sadness, #unhappy, #grief), *surprise* (#surprise, #surprised), *love* (#love), *shame* (#shame), *anger* (*e.g.* #anger, #rage, #hate), *fear* (*e.g.*, #fear, #scare, #worry), *disgust* (#disgust), *trust* (#trust). From this overall set $\mathcal{D}$, we subsample three corpora, each with 200,000 instances, such that empirical testing of estimated models is not affected by different training set sizes. Sizes for these sets are shown in Table 1. From these subcorpora, we use 150,000 randomly sampled instances for training and 50,000 for testing.

The corpus D_{all} is sampled without any constraints. It therefore contains posts with and without image attachments. The subcorpus D_{OCR} is sampled from all instances for which the optical character recognition generated output longer than

Table 2: Results for Tweet text features alone, on different sets (Experiment 1) and for OCR features (ϕ^d_{OCR}) and caption features (ϕ^d_{Vis}) in isolation (Experiment 2).

	Experiment 1									Experiment 2					
	ϕ^d_{text}									ϕ^d_{OCR}			ϕ^d_{Vis}		
	D_{all}			D_{OCR}			D_{Vis}			D_{OCR}			D_{Vis}		
Emotion	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
joy	78	86	82	81	89	85	79	88	84	70	83	76	57	98	72
love	70	72	71	75	78	77	73	73	73	62	58	60	47	6	11
sadness	75	72	73	80	69	74	76	56	65	58	46	51	58	1	2
fear	82	70	75	86	76	81	79	60	68	84	70	76	18	0	0
anger	84	68	75	86	70	77	80	50	61	80	62	69	91	9	17
shame	89	65	75	86	52	65	63	17	27	56	40	47	0	0	0
surprise	75	48	58	65	32	43	62	31	42	53	20	29	0	0	0
trust	85	65	74	84	65	73	84	59	69	76	60	67	45	2	4
disgust	73	39	51	17	4	6	0	0	0	0	0	0	0	0	0
Macro-Av.	79	65	71	73	59	65	66	48	54	60	49	53	35	13	12

six bytes (*i.e.*, $\forall d \in D_{\text{OCR}} : \phi^d_{\text{text}} \neq \varnothing$). The subcorpus D_{Vis} is sampled from all instances for which an image was linked, but for which the optical character recognition did *not* return any output (*i.e.*, $\forall d \in D_{\text{Vis}} : \phi^d_{\text{Vis}} \neq \varnothing \wedge \phi^d_{\text{OCR}} = \varnothing$). We therefore assume that D_{OCR} consists of Tweets with images which contain text and D_{Vis} consists of Tweets with images without text.

The first three columns of Table 1 show the numbers of Tweets for each emotion, using those which contain ϕ^d_{OCR} and ϕ^d_{Vis} features, respectively. Interestingly, the amount of Tweets with text related to *fear* and *trust* is higher than for other images. Tweets which are associated with *love*, *joy*, *fear* and *trust* contain more images without text than other emotions. In general, the portion of Tweets with emotion hashtags including images is 18 %.

3 Results

In the following, we discuss three experiments: In Experiment 1, we analyze Tweet features only (ϕ^d_{text}), but on three different sets (D_{all}, D_{OCR}, D_{Vis}). The left part of Table 2 shows the results as F_1, precision and recall. We observe that, on average, Tweet text features ϕ^d_{text} are sufficient to predict emotion labels, also for Tweets which contain an image with text: The performance of ϕ^d_{text} is not dramatically different on average for sets D_{all} and D_{OCR} (when the very infrequent emotion classes, especially *disgust*, are not taken into account). However, performance improves for *joy*, *love*, *sadness*, *fear*, and *anger*, while it decreases for the others. For Tweets with images without embedded text (D_{Vis}), text features are not as sufficient; we observe a clear drop in most emotions (except of *joy* amd *love*).

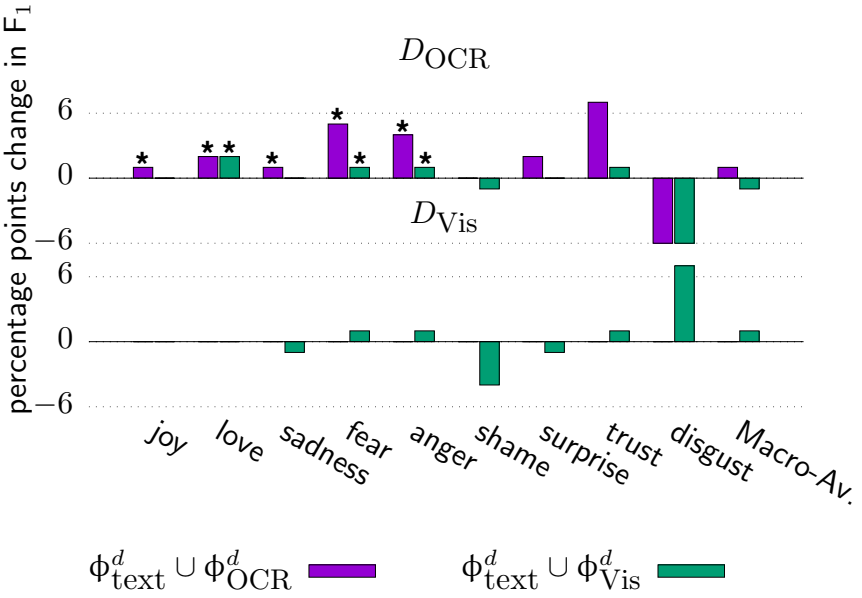

Fig. 1: Experiment 3: Complementarity of ϕ^d_{OCR} and ϕ^d_{Vis} to ϕ^d_{text}. The baseline corresponds to results with ϕ^d_{text} only as shown in Table 2. Significant differences ($\alpha = 0.01$) are denoted with a * (tested via bootstrap resampling [18]).

In Experiment 2 (also shown in Table 2), we analyze the classification performance without ϕ^d_{text}, but with ϕ^d_{OCR} alone (on ϕ^d_{OCR}) and with ϕ^d_{Vis} alone (on ϕ^d_{Vis}). For, ϕ^d_{OCR}, the performance is still good, but lower than for ϕ^d_{text}, therefore though a text image is added to the Tweet, the text of Tweet is a more important signal than the text in the recognized text in the image. The features ϕ^d_{Vis} are not sufficient for acceptable classification performance.

In Experiment 3, we analyze if ϕ^d_{OCR} or ϕ^d_{Vis} complement the information in ϕ^d_{text} in D_{OCR} and D_{Vis}. Figure 1 shows these differences; the baseline corresponds to results in Table 2. Though ϕ^d_{Vis} is not sufficient for classification alone, it contributes slightly to *fear*, *anger*, *trust*, and *disgust* on D_{Vis}, however, the positive impact is limited. The contribution of ϕ^d_{OCR} in addition ϕ^d_{text} on D_{OCR} is substantial, with up to 7 percentage points in F_1 for *trust*, 5pp for *fear* and 4pp for *anger*.

4 Conclusion & Future Work

In this paper, we investigated the effect of features from optical character recognition and caption generation on emotion classification from Tweets. While the caption generation does only help generalization in few cases (in which a qualitative analysis shows an actual better generalization across image content), the OCR information is of clear importance. Therefore we can conclude that textual information from images contributes a helpful signal which complements text features from Tweets. This contribution is especially large for *fear*, *anger* and *trust*. Interesting are *shame* and *surprise*, for which classification on Tweets

with images which contain text is more difficult based on Tweet text features alone. This drop can only be compensated partially with OCR.

Generated captions could not be shown to contribute substantially; one possible reason is that the generated text is too abstract. Future work will therefore focus on features from intermediate levels of the deep neural network.

References

1. Ekman, P.: Basic emotions. In Dalgleish, T; Power, M., ed.: Handbook of Cognition and Emotion. John Wiley & Sons, Sussex, UK (1999)
2. Plutchik, R.: The nature of emotions. American Scientist (2001)
3. Alm, C.O., Roth, D., Sproat, R.: Emotions from text: Machine learning for text-based emotion prediction. In: HLT-EMNLP. (2005)
4. Aman, S., Szpakowicz, S.: Identifying expressions of emotion in text. In: TSD. (2007)
5. Dodds, P.S., Harris, K.D., Kloumann, I.M., Bliss, C.A., Danforth, C.M.: Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. PloS one **6**(12) (2011)
6. Costa, J., Silva, C., Antunes, M., Ribeiro, B.: Concept drift awareness in twitter streams. In: ICMLA. (2014)
7. Wang, W., Chen, L., Thirunarayan, K., Sheth, A.P.: Harnessing twitter "big data" for automatic emotion identification. In: SocialCom/PASSAT. (2012)
8. @kezia_hunter: Be prepared for a.... Twitter (2017) https://twitter.com/kezia_hunter/status/818260781108228098.
9. Bartlett, M.S., Littlewort, G., Lainscsek, C., Fasel, I., Movellan, J.: Machine learning methods for fully automatic recognition of facial expressions and facial actions. In: SMC. (2004)
10. Kahou, S.E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, c., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R.C., Mirza, M., Jean, S., Carrier, P.L., Dauphin, Y., Boulanger-Lewandowski, N., Aggarwal, A., Zumer, J., Lamblin, P., Raymond, J.P., Desjardins, G., Pascanu, R., Warde-Farley, D., Torabi, A., Sharma, A., Bengio, E., Côté, M., Konda, K.R., Wu, Z.: Combining modality specific deep neural networks for emotion recognition in video. In: ICMI. (2013)
11. Gupta, A., Lamba, H., Kumaraguru, P., Joshi, A.: Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy. In: WWW. (2013)
12. Wang, Y., Wang, S., Tang, J., Liu, H., Li, B.: Unsupervised sentiment analysis for social media images. In: IJCAI. (2015)
13. Smith, R., Inc, G.: An overview of the tesseract ocr engine. In: ICDAR. (2007)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012)
15. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR. (2015)
16. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR. (2015)
17. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. CoRR **abs/1405.0312** (2014)
18. Efron, B.: Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics **7**(1) (1979) 1–26