Modeling Emoji Generation for Emotion Analysis of Social Media Short Texts

Sujatha Das Gollapalli, See-Kiong Ng
Institute of Data Science, National University of Singapore, Singapore
{idssdg, seekiong}@nus.edu.sg

Abstract

Emojis and emoticons are widely employed in user-generated content on social media. Existing generative models for short texts do not particularly handle emojis and emoticons thereby missing the extra information conveyed by these "special" expressions. We present *EmDMM*, a novel Dirichlet Multinomial Mixture model for capturing emotions expressed through emojis and emoticons in social media short texts. *EmDMM* can automatically detect emoji clusters that reflect emotion classes providing an unsupervised tool for analyzing emotions in rapidly-emerging social media content. We apply *EmDMM* to COVID-19 tweets and extract public emotions on topics related to the on-going pandemic.

Motivation

Capturing sentiments and emotions in user-generated content is useful for various applications in e-commerce, public health, and disaster management (Bollen, Mao, and Zeng 2011; Hu et al. 2018). On social media platforms, many users add emojis[1], emoticons, and slang abbreviations (such as "LOL/SMH") for conveying non-verbal cues, emphasis, and tone. Although emojis can serve various functions in interpersonal communications for complementing and condensing messages (e.g. "Having ☕", 🤙), face emojis (e.g. 😄) and emoticons (e.g. **:-)**) that specifically mirror facial expressions in written text are used to express emotions (Kelly and Watts 2015; Hogenboom et al. 2015; Guibon, Ochs, and Bellot 2018).

Various studies have examined the semantics and sentiment expressed by emojis (Barbieri, Ronzano, and Saggion 2016; Novak et al. 2015). These studies note that emojis and emoticons provide a "common language" for emotive expression across several languages, cultures, and platforms (Barbieri et al. 2016; Ljubešić and Fišer 2016; Li et al. 2019). Consequently, emojis were found to provide additional cues in sentiment, emotion and sarcasm detection, as well as user-personality modeling tasks (Davidov, Tsur, and Rappoport 2010; Li et al. 2018; Hussien et al. 2019).

Some recent studies address the usage, semantics, and predictive power of emojis. For example, similar to word embeddings, emoji embeddings were analyzed for modeling the similarities, differences, and clustering of emojis (Barbieri, Ronzano, and Saggion 2016; Eisner et al. 2016; Felbo et al. 2017) whereas emoji recommendation and prediction was addressed as part of SemEval tasks[2] in Natural Language Processing research (Barbieri et al. 2018).

However, to the best of our knowledge, none of the existing studies on emojis address the generative process behind emojis. Just as word generation depends on a *latent* topic (Blei, Ng, and Jordan 2003), we argue that generation of emojis (specifically, face and gesture emojis and emoticons) depends on a *latent* emotion and modeling this aspect provides a more complete understanding of user-generated content. For the rest of our paper, unless otherwise indicated, we use the term *emojis* to jointly refer to both "emojis" and "emoticons" and the terms "document/tweet/short text" are used interchangeably.

Contributions: We present *EmDMM*, our extension to the Dirichlet Multinomial Mixture topic model commonly used for modeling short texts. In *EmDMM*, we capture token generation in short texts through two latent variables, one representing the *topic* (as in standard topic models) and the second, representing an *emotion*. We show that *EmDMM* effectively captures emoji clusterings indicative of the underlying emotions through an analysis of recently-collected tweets related to the coronavirus pandemic. Our analysis showcases *EmDMM*'s novel capability to track emotions in rapidly-emerging social media short texts in an unsupervised manner.

EmDMM for modeling Emojis

Background: The document generative process in Dirichlet Multinomial Models (DMM) is based on the assumption that a given corpus can be viewed as a mixture of latent topics (McCallum and Nigam 1998). Given a set of K topics, during document generation, first, a topic is chosen from this set and each observed word in the document is conditioned on the chosen topic. While this assumption was found to be restrictive for long documents (Blei, Ng, and Jordan 2003), due to their sparse nature, short texts are often modeled using DMMs (Zhao et al. 2011; Yin and Wang 2014;

[1] http://emojitracker.com/

[2] https://competitions.codalab.org/competitions/17344

Li, Zhang, and Ouyang 2019).

We adopt the modeling construct of DMMs for short texts and extend it with the additional capability to view the corpus along two dimensions namely, a mixture of K topics ($\vec{\theta_t}$) as well as a mixture of E emotions ($\vec{\theta_e}$). In *EmDMM*, given a short text with emojis, (1) the list of words is conditioned on a *latent* topic z, and (2) the list of emojis is conditioned on a *latent* emotion u. The generative process for the corpus in *EmDMM* along with the graphic model is shown in Figure 1. $\alpha_t, \alpha_e, \beta_t, \beta_e$ are Dirichlet hyperparameters while Dir and $Mult$ refer to Dirichlet and Multinomial distributions, respectively (Heinrich 2004).

Due to space limits, we refer the reader to (Yin and Wang 2014) for the derivation of sampling equations for DMMs. Since words and emojis are independently modeled in *EmDMM*, the Gibbs sampling equations for variables corresponding to latent emotions (u) mirror those for latent topics (z) except that they operate on emotion mixtures ($\vec{\theta_e}$) and "observed words" come from the emoji alphabet. Our sampling equations and variable explanations are listed in Table 1. DMM provides a natural clustering of documents due to the single topic per single document assignments (McCallum and Nigam 1998). Likewise, with *EmDMM*, each document is assigned a "topic cluster label" as well as an "emoji cluster label" resulting in a natural clustering of documents along these two dimensions.

COVID-19 Tweet Analysis

We use *EmDMM* to examine the *covid19_twitter* dataset provided by Banda, et al (2020).[3] Approximately 100K English tweet ids were sampled per week and processed using the Social Media Mining Toolkit.[4] We removed the keywords used in data gathering such as "coronavirus, CoronavirusPandemic, COVID-19" and employed corpus frequency thresholds for collecting the term, hashtag, and emoji dictionaries (20, 5, and 3, respectively). The dictionary sizes are emojis/emoticons:1,318, hashtags:22,772, and terms:43,138. Emojis and emoticons were extracted using the *emojificate* and *emot* libraries available in Python. All text content was normalized by removing punctuation, stopwords, and non-alphanumeric tokens and converting all remaining tokens to lowercase. Our final dataset for *EmDMM* runs contains a total of 333,937 tweets with emojis over the period: March 01, 2020 to Sept 13, 2020.

EmDMM was implemented in Java by extending the topic models provided in the Mallet toolkit (McCallum 2002).[5] As in previous works, $\alpha_t, \alpha_e, \beta_t, \beta_e$ parameters are set to a small value (0.1) and the number of iterations for Gibbs sampling set to 1000 (Yin and Wang 2014). All *EmDMM* runs took approximately 1-2 hours depending on the number of topic and emoji clusters on an Intel Xeon 2.90GHz CPU machine with 32GB RAM.

Number of Topic/Emoji Clusters: Based on previous studies on hashtag usage for labeling and clustering tweets (Mehrotra et al. 2013; Kunneman, Liebrecht, and van den Bosch 2014), we used hashtags as "soft" labels to set the number of topic clusters using the cluster purity measure (Aggarwal and Zhai 2012). Since the cluster purity increases with increasing number of clusters and we want to avoid too many very small clusters, we trade-off between these two factors and experimented in the range $\{10, 20 \ldots 100\}$ for the number of topic clusters. We observed that for settings above 60, more than 50% of resulting clusters cover less than 0.1% of the data points. Based on previous studies in Psychology that proposed a small set of $6 - 8$ basic emotions (Plutchik 2001; Ekman 1992) and to account for other emoji groups (such as "objects/countries"), we tested the number of emoji clusters from the set $\{6, 8, 10, 12\}$ and choose 12 based on the emoji cluster name labeling experiment (described next). All further analysis is based on *EmDMM* output with 60 topic clusters and 12 emoji clusters.

Emoji Cluster Names: To obtain the latent emotion captured by a specific emoji cluster extracted by *EmDMM*, we used the "SemEval-18 Affect in Tweets" dataset. To our knowledge, this is the largest dataset with emotion-labeled tweets that also have a reasonable proportion (about ten percent) of tweets with emojis (Bostan and Klinger 2018). The tweets were labeled using Plutchik's basic emotions: {*joy, trust, anticipation, surprise, sadness, fear, anger, and disgust*} (Mohammad et al. 2018). Note that in this list, the first four emotions comprise "positive" emotions whereas the last four are negative emotions. We computed the emoji-emotion probabilities using co-occurrence counts based on tweet-level labels in this dataset and calculated the aggregate probability of top-5 emojis/emoticons for a cluster with respect to each emotion from the Plutchik's set. An emoji cluster is assigned the best matching emotion name, if and only if, if its score is thrice as high as its closest emotion from the opposite class. We adopted this stringent matching since we compute emoji-emotion probabilities based on a small labeled dataset and the same emoji can be used with opposing emotions (Novak et al. 2015). Our procedure resulted in unique emotion names for 4 of the 12 emoji clusters shown in Table 2 unlike other settings of number of emoji clusters.

As seen in Table 2, both the lists of clustered emojis as well as the assigned emotion labels are both intuitive and consistent (Guibon, Ochs, and Bellot 2018). We observe that non-facial and non-gesture emojis serve purposes such as marking geographical location through regional indicator emojis (in case of clusters 8 and 9). Emojis in clusters 4 and 5 seem to express "approval" and "ridicule", respectively, but do not correspond to the basic emotions (Plutchik 2001).

Corpus Composition: The cluster sizes produced by *EmDMM* is skewed along both topic/emoji dimensions. The top-3 (5%) topics cover almost 44.84% of the corpus whereas the bottom-30% topics are assigned to a mere 1.3% of tweets in the dataset. The percentage of tweets assigned to the emoji clusters range from $2 - 18\%$ with the top-3 emoji clusters covering about 41.6% of tweets. In particular, the emoji cluster ids 6 and 11 corresponding to "positive" emotion (*joy*) are assigned to about 26.3% of the tweets whereas cluster ids 3 and 10 corresponding to "negative"

[3]https://github.com/thepanacealab/covid19_twitter version 27, collected on Sept 19, 2020

[4]https://github.com/thepanacealab/SMMT

[5]http://mallet.cs.umass.edu/

1. Draw $\vec{\theta_e} \sim Dir(\alpha_e), \vec{\theta_t} \sim Dir(\alpha_t)$
2. For each latent topic $k = 1 \ldots K$,
   draw $\vec{\phi_t^k} \sim Dir(\beta_t)$
   For each latent emotion $u = 1 \ldots E$,
   draw $\vec{\phi_e^u} \sim Dir(\beta_e)$
3. For each tweet d in the corpus
   (i) Draw $z \sim Mult(\vec{\theta_t}), u \sim Mult(\vec{\theta_e})$
   (ii) For each word w in d, draw $w \sim Mult(\vec{\phi_t^z})$
   (iii) For each emoji m in d, draw $m \sim Mult(\vec{\phi_e^u})$

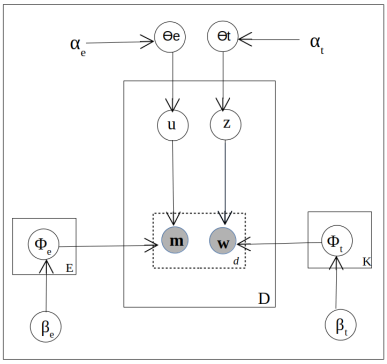

Figure 1: The generative process in *EmDMM* is shown along with the plate diagram illustrating the graphical model.

D	number of documents in corpus		
K	number of topic-clusters	E	number of emoji-clusters
V_w	number of words in word-vocabulary	V_e	number of emojis in emoji-vocabulary
$\vec{z}$	topic-cluster labels of each document	$\vec{e}$	emoji-cluster labels of each document
p_z	number of documents in topic-cluster z	p_e	number of documents in emoji-cluster e
n_z	number of words in topic-cluster z	n_e	number of emojis in emoji-cluster e
n_z^w	number of occurrences of word w in topic-cluster z	n_e^m	number of occurrences of emoji m in emoji-cluster e
N_d	number of words in document d	N_m	number of emojis in document d
N_d^w	Number of occurrences of word w in d	N_d^m	Number of occurrences of emoji m in d

$$p(z_d = z|z_{\neg d}, \mathbf{D}) \propto \left(\frac{p_{z,\neg d}+\alpha_t}{D-1+K\alpha_t}\right)\left(\frac{\prod_{w\in d}\prod_{j=1}^{N_d^w}(n_{z,\neg d}^w+\beta_t+j-1)}{\prod_{i=1}^{N_d}(n_{z,\neg d}+V_w\beta_t+i-1)}\right)$$

$$p(e_d = e|e_{\neg d}, \mathbf{D}) \propto \left(\frac{p_{e,\neg d}+\alpha_e}{D-1+E\alpha_e}\right)\left(\frac{\prod_{m\in d}\prod_{j=1}^{N_d^m}(n_{e,\neg d}^m+\beta_e+j-1)}{\prod_{i=1}^{N_m}(n_{e,\neg d}+V_e\beta_e+i-1)}\right)$$

Table 1: Summary of variables used in *EmDMM* equations

EID	Emojis	EID	Emojis
0	😡😈😱⭐💙	1	👇💀👉👀
2	➖🚨➡️ ▪ ✅	4	👏☠️👍⬇️🖕
5	♂️♀️💪👱👩	7	😂🤣🤔😅BPB*
8	🦠NARI	9	SU🔥E🙏

EID	Emojis	Emotion Labels
3	😒🙄💩😷😢	Disgust
6	❤️✅🙏😍d:	Joy
10	😭💔😢🏈⚠️	Sadness
11	🤣🙏🤡😂💞	Joy

Table 2: Emoji clusters extracted by *EmDMM*. *BPB stands for "Braille Pattern Blank". EID refers to the emoji-cluster id and the clusters mapped to specific emotions are highlighted separately.

Topic	Top-Hashtags
45 (19%)	#coronavirussa #covididiots #coronapocalypse #riots #covidiot #coronavirusoutbreak
20 (14%)	#fakenews #dumptrump #ccp #theresistance #democrats #fauci #who #billgates
35 (11%)	#savetheworld #billionshields #quarantine #school #quarantinelife #backtoschool
42 (7%)	#africa #blog #leadership #innovation #technology #publichealth #healthcare
32 (6%)	#coronaviruschina #wuhan #covid– #thankyou #coronapocalypse

Table 3: Top topics based on corpus coverage and their hashtags

emotions (*disgust/sadness*) are assigned to about 17.9% of the dataset. These high percentages support previous studies which noted that social media platforms such as Twitter enable not only the exchange of urgent and critical information but also serve as outlets for emotion expression during disasters and calamities (Ashktorab et al. 2014).

In Table 3, we show the top-5 topics based on the percentage of corpus assigned to the topic along with their top hashtags (based on the exclusivity measure (McCallum 2002)). Not surprisingly, the topics capture content clustering along various themes related to the pandemic such as *quarantine, healthcare,* and *WHO* along with on-going events such as *resumption of schools* and *elections*. About 8-16% of tweets from these topics are assigned to positive and negative emotions as shown in Figure 2. For all topics shown here, the percentage of tweets assigned to the "positive" clusters is higher than that for the "negative" clusters indicating that in pandemic tweets, emojis were used more frequently while expressing "positive" sentiment than while expressing "negative" sentiment.

Emotion Detection: Since large-scale, emotion-labeled datasets for tweet with emojis are missing (Bostan and Klinger 2018), we illustrate the emotion detection capability of *EmDMM* using anecdotal tweets assigned to emotion clusters from the top-3 topics in Table 4. In these tweets, we note that Twitter users express emotions on aspects of their personal lives affected by the pandemic using emojis and text. The emotion cluster labels assigned

TID	*EmDMM*	DepecheMood	ESTeR	Tweet
45	Joy	Happy	Anticipation	Come rain, heat, snow or coronavirus the bank's demands make it through my letter box. :)
45	Disgust	Afraid	`Joy`	Couldn't agree more. This is a big fuck you from Mother Nature herself 🙌🟫
45	Sadness	Afraid	`Surprise`	Shit..... Hope he gets well soon 🙏🙏
20	Joy	Inspired	Joy	That's Humanity ❤️#CoronaVirus #COVID
20	Disgust	`Inspired`	`Neutral`	Thanks coronavirus for ruining ALL of my plans 🙃
20	Sadness	`Inspired`	`Joy`	God!! Something more coming to us!! 💔💔💔
35	Joy	`Annoyed`	Joy	Share with the boys in your life! 🙌
35	Disgust	`Inspired`	Fear	🥺🥺🥺reasons i am not flying anywhere
35	Sadness	`Happy`	Disgust	New Coronavirus definition: Something that is utilized to take away everything that makes life bearable. 😒

Table 4: Anecdotal tweets and predicted emotion labels for the top-3 topics 45, 20, and 35. The `highlighted` predicted labels were judged unsuitable by us for the given tweet content.

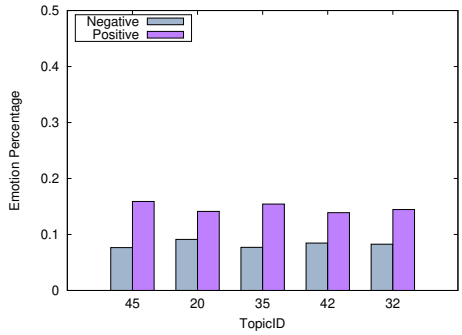

Figure 2: Emotion spread in tweets of top-5 topics

by *EmDMM* are shown along with labels predicted by two state-of-the-art unsupervised emotion detection models: DepecheMood (Araque et al. 2019) and ESTeR (Gollapalli, Rozenshtein, and Ng 2020).

The DepecheMood lexicon contains about $188K$ words whose scores with respect to emotions: {*'afraid', 'amused', 'angry', 'annoyed', 'dontcare', 'happy', 'inspired', 'sad'*} were compiled using distant supervision (Araque et al. 2019). Each tweet is assigned a score for an emotion by aggregating scores of words comprising the tweet. In contrast, ESTeR assigns emotion labels from Plutchik's set of basic emotions: {*joy, trust, anticipation, surprise, sadness, fear, anger, and disgust*} or *neutral* using emotion-sensitive PageRank on a word graph built from a large corpus such as Wikipedia (Gollapalli, Rozenshtein, and Ng 2020).

As such, since the three models (*EmDMM*, ESTeR, DepecheMood) operate on different label sets, they are not directly comparable. Still as anecdotes, we show the emotion labels produced by *EmDMM* for the examples in Table 4. We specifically chose examples with sparse or ambiguous textual content where emojis provide valuable cues for identifying the emotion being expressed that are harnessed by our model (for example, the tweets for topic 35).

Since we modeled a COVID-specific tweet corpus using *EmDMM*, and do not have labeled tweets with emojis to evaluate *EmDMM* emotion labels quantitatively, for an anecdotal evaluation, we randomly sampled about 30 tweets each for "Joy/Sadness/Disgust" from the top-3 topic clusters produced by *EmDMM* and manually evaluated the assigned emotion labels. The manually-assigned label counts were {*joy*:20, *sadness*:25, *disgust*:34, *Other*:11} and *EmDMM*'s accuracy was 48%. Though this value is not very high in absolute terms, we note that the number is significantly better than a random baseline and ours is notably an unsupervised model making it applicable for emotion analysis of the emerging content on social media portals.

Related Work

Short texts (such as tweets) have unique characteristics due to their sparsity, noise (typos and misspellings), as well as additional lingusitic features such as #hashtags. Various extensions to the basic LDA(Blei, Ng, and Jordan 2003) were therefore developed for tweets. These extensions incorporate word co-occurrences via bigrams, tweet pooling and aggregation (Yan et al. 2013; Yin and Wang 2014; Chen et al. 2015), user-specific topic mixtures (Zhao et al. 2011), as well as temporal dynamics (Sasaki, Yoshikawa, and Furuhashi 2014). Topic models were also studied for analyzing opinion, sentiments and social emotions in user-generated content (Bao et al. 2011; Lim and Buntine 2014; Nguyen and Shirai 2015; Quan et al. 2015). We refer the interested reader to Bostan and Klinger (2018) for an overview on emotion recognition.

Conclusions

We developed *EmDMM*, a topic model that incorporates emojis into the generation process of short texts and enables corpus analysis along two dimensions: topics and emotions. By harnessing emojis, *EmDMM* is able to provide an unsupervised tool for emotion analysis of the rapidly-emerging social media content. *EmDMM* presents a first-cut extension to DMMs for incorporating emojis into the short-text generation process. For future work, we would like to explore joint models that capture the dependence between emojis and words, and study how to handle user-specific and temporal variables, as well as content containing mis-

spellings, slang, and abbreviations (Zhao et al. 2011; Sasaki, Yoshikawa, and Furuhashi 2014).

References

Aggarwal, C. C.; and Zhai, C. 2012. A survey of text clustering algorithms. In *Mining text data*.

Araque, O.; Gatti, L.; Staiano, J.; and Guerini, M. 2019. Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques. *IEEE transactions on affective computing* .

Ashktorab, Z.; Brown, C.; Nandi, M.; and Culotta, A. 2014. Tweedr: Mining twitter to inform disaster response. In *ISCRAM*.

Banda, J. M.; Tekumalla, R.; Wang, G.; Yu, J.; Liu, T.; Ding, Y.; Artemova, K.; Tutubalina, E.; and Chowell, G. 2020. A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration.

Bao, S.; Xu, S.; Zhang, L.; Yan, R.; Su, Z.; Han, D.; and Yu, Y. 2011. Mining social emotions from affective text. *IEEE TKDE* .

Barbieri, F.; Ballesteros, M.; Ronzano, F.; and Saggion, H. 2018. Multimodal Emoji Prediction. In *NAACL-HLT*.

Barbieri, F.; Kruszewski, G.; Ronzano, F.; and Saggion, H. 2016. How Cosmopolitan Are Emojis? Exploring Emojis Usage and Meaning over Different Languages with Distributional Semantics. In *International Conference on Multimedia*.

Barbieri, F.; Ronzano, F.; and Saggion, H. 2016. What does this Emoji Mean? A Vector Space Skip-Gram Model for Twitter Emojis. In *LREC*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* .

Bollen, J.; Mao, H.; and Zeng, X. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* .

Bostan, L.-A.-M.; and Klinger, R. 2018. An Analysis of Annotated Corpora for Emotion Classification in Text. In *COLING*.

Chen, W.; Wang, J.; Zhang, Y.; Yan, H.; and Li, X. 2015. User Based Aggregation for Biterm Topic Model. In *ACL-IJCNLP*.

Davidov, D.; Tsur, O.; and Rappoport, A. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *COLING*.

Eisner, B.; Rocktäschel, T.; Augenstein, I.; Bošnjak, M.; and Riedel, S. 2016. emoji2vec: Learning Emoji Representations from their Description. In *Workshop on Natural Language Processing for Social Media*.

Ekman, P. 1992. An argument for basic emotions. *Cognition & emotion* 169–200.

Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; and Lehmann, S. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *EMNLP*.

Gollapalli, S. D.; Rozenshtein, P.; and Ng, S.-K. 2020. ESTeR: Combining Word Co-occurrences and Word Associations for Unsupervised Emotion Detection. In *Findings of EMNLP*.

Guibon, G.; Ochs, M.; and Bellot, P. 2018. From Emoji Usage to Categorical Emoji Prediction. In *COLING 2018*.

Heinrich, G. 2004. Parameter Estimation for Text Analysis. Technical report. URL http://www.arbylon.net/publications/text-est.pdf.

Hogenboom, A.; Bal, D.; Frasincar, F.; Bal, M.; De Jong, F.; and Kaymak, U. 2015. Exploiting Emoticons in Polarity Classification of Text. *J. Web Eng.* .

Hu, T.; Xu, A.; Liu, Z.; You, Q.; Guo, Y.; Sinha, V.; Luo, J.; and Akkiraju, R. 2018. Touch Your Heart: A Tone-Aware Chatbot for Customer Care on Social Media. In *CHI*.

Hussien, W.; Al-Ayyoub, M.; Tashtoush, Y.; and Al-Kabi, M. 2019. On the Use of Emojis to Train Emotion Classifiers. *arXiv preprint arXiv:1902.08906* .

Kelly, R.; and Watts, L. 2015. Characterising the inventive appropriation of emoji as relationally meaningful in mediated close personal relationships. In *Experiences of Technology Appropriation: Unanticipated Users, Usage, Circumstances, and Design*.

Kunneman, F.; Liebrecht, C.; and van den Bosch, A. 2014. The (Un)Predictability of Emotional Hashtags in Twitter. In *Workshop on Language Analysis for Social Media (LASM)*.

Li, M.; Guntuku, S.; Jakhetiya, V.; and Ungar, L. 2019. Exploring (Dis-)Similarities in Emoji-Emotion Association on Twitter and Weibo. In *WWW*.

Li, W.; Chen, Y.; Hu, T.; and Luo, J. 2018. Mining the relationship between emoji usage patterns and personality. In *ICWSM*.

Li, X.; Zhang, J.; and Ouyang, J. 2019. Dirichlet Multinomial Mixture with Variational Manifold Regularization: Topic Modeling over Short Texts. In *AAAI*.

Lim, K. W.; and Buntine, W. 2014. Twitter Opinion Topic Model. *CIKM* .

Ljubešić, N.; and Fišer, D. 2016. A Global Analysis of Emoji Usage. In *Proceedings of the 10th Web as Corpus Workshop*.

McCallum, A.; and Nigam, K. 1998. A Comparison of Event Models for Naive Bayes Text Classification. In *AAAI Workshop on Learning for Text Categorization*, 41–48.

McCallum, A. K. 2002. MALLET: A Machine Learning for Language Toolkit. Http://mallet.cs.umass.edu.

Mehrotra, R.; Sanner, S.; Buntine, W.; and Xie, L. 2013. Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling. In *SIGIR*.

Mohammad, S. M.; Bravo-Marquez, F.; Salameh, M.; and Kiritchenko, S. 2018. SemEval-2018 Task 1: Affect in Tweets. In *SemEval-2018*.

Nguyen, T. H.; and Shirai, K. 2015. Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction. In *ACL-IJCNLP*.

Novak, P. K.; Smailovic, J.; Sluban, B.; and Mozetic, I. 2015. Sentiment of Emojis.

Plutchik, R. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist* 344–350.

Quan, X.; Wang, Q.; Zhang, Y.; Si, L.; and Wenyin, L. 2015. Latent Discriminative Models for Social Emotion Detection with Emotional Dependency. *ACM Trans. Inf. Syst.* .

Sasaki, K.; Yoshikawa, T.; and Furuhashi, T. 2014. Online topic model for Twitter considering dynamics of user interests and topic trends. In *EMNLP*.

Yan, X.; Guo, J.; Lan, Y.; and Cheng, X. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*.

Yin, J.; and Wang, J. 2014. A Dirichlet Multinomial Mixture Model-Based Approach for Short Text Clustering. In *KDD*.

Zhao, W. X.; Jiang, J.; Weng, J.; He, J.; Lim, E.-P.; Yan, H.; and Li, X. 2011. Comparing Twitter and Traditional Media Using Topic Models. In *ECIR*.