CTM - A Model for Large-Scale Multi-View Tweet Topic Classification

Anonymous ACL submission

Abstract

Automatically associating social media posts with topics is an important prerequisite for effective search and recommendation on many social media platforms. However, topic classification of such posts is quite challenging because of (a) a large topic space (b) short text with weak topical cues and (c) multiple topic associations per post. In contrast to most prior work which only focuses on post classification into a small number of topics ($10-20$), we consider the task of large-scale topic classification in the context of Twitter where the topic space is 10 times larger with potentially multiple topic associations per Tweet. We address the challenges above and propose a novel neural model, CTM that (a) associates tweets from a large topic space of 300 topics (b) takes a holistic approach to tweet content modeling – leveraging multi-modal content, author context, and deeper semantic cues in the Tweet. We evaluate CTM quantitatively and show that our method offers an effective way to classify Tweets into topics at scale and is superior in performance to other approaches yielding a significant relative lift of $\mathbf{20}\%$.

1 Introduction

On many social media platforms like Twitter, users find posts that they are interested in through two mechanisms: (a) search and (b) recommendation. Both mechanisms typically use the topics associated with posts to identify potential candidates that are displayed to the user. Therefore, automatically associating a post with topics is important for effective search and recommendation. Furthermore, due to the diverse nature of social media content, for such topic association to be useful in practice, it is important to (a) support classification into a large number of topics (potentially hundreds or thousands of topics) and (b) allow for a post to have multiple topics or no topic at all.

While traditionally, there has been a long line of work on classifying documents (like news articles, movie reviews etc.) into topics spanning half a century (Borko and Bernick, 1963; Balabanovic and Shoham, 1995; Joachims, 1998; Tsutsumi et al., 2007; Yang et al., 2014; Adhikari et al., 2019), classifying social media content poses several unique challenges (Chang et al., 2015). First, such posts can be very short (at-most 280 characters on Twitter) and noisy where cues provided by the linguistic context alone can be very sparse (Baldwin et al., 2013). Second, social media content is multi-modal with associated images, videos, and hyperlinks. Approaches for classifying documents tend to ignore this multi-modal nature (Chang et al., 2015). With the rise of social media platforms, several works do explore classification of social media posts (like Tweets) (Lee et al., 2011; Genc et al., 2011; Tao et al., 2012; Stavrianou et al., 2014; Selvaperumal and Suruliandi, 2014; Cordobés et al., 2014; Kataria and Agarwal, 2015; Chang et al., 2015; Li et al., 2016b,c,d; Ive et al., 2018; Kang et al., 2019; Gonzalez et al., 2021). However, all of these works suffer from one or more limitations: (a) Only support a small set of topics (about 10 topics or categories) (b) model only the text, ignore multi-modal content, deeper semantic-cues and (c) do not support multiple labels per post.

In this paper, we address the above challenges in the context of Tweet classification. We propose CTM, a Tweet topic classification model that (a) supports classification into 300 topics (10 times larger than prior work) (b) incorporates rich content like media, hyperlinks, author features, entity features thus moving beyond shallow Tweet text features and (c) supports multiple topics to be associated per Tweet. First, we construct a moderate-sized high-quality human annotated labeled dataset and a large dataset of weakly labeled examples to use for fitting our predictive model. We then propose a neural model that models a Tweet holistically (including text, media, hyperlinks, entities etc.) to annotate Tweets with topics. In addition to modeling several aspects of a Tweet, we also encode specific label constraints in a principled manner using probabilistic inference over a factor graph. Our method offers an effective way to classify Tweets into topics at scale and is superior in performance to other approaches yielding a significant relative lift of **20**%. More broadly, our proposed model also reinforces the central role of larger contextual cues in the predictive modeling of social media content – the work-horse powering many social media platforms.

2 Related Work

There is a long line of work on topic classification of Tweets spanning more than a decade since the inception of Twitter. Early works used bag-of-words features extracted from Tweet text to classify Tweets into topics using standard classifiers like Rocchio classifiers, logistic regression, and support-vector machines (Lee et al., 2011; Genc et al., 2011; Tao et al., 2012; Stavrianou et al., 2014; Selvaperumal and Suruliandi, 2014). All of these works focus on just using the Tweet text with accompanying hashtags and consider only a small number of topics.

Following this initial line of work, some works investigated using increasingly rich features for topic classification. Cordobés et al. (2014) used graph-based metrics like page-rank on term co-occurrence graphs to predict topics while Kataria and Agarwal (2015) used hyperlink information in addition to Tweet text. One of the first works to consider modeling content beyond raw text of social media was the work of (Chang et al., 2015). In particular, Chang et al. (2015) explored the problem of classifying Google+ posts into topics. Noting that Google+ posts may include media, they investigated the problem of ensembling predictions of independent black-box models, each making topic predictions from a different modality. They primarily show how leveraging crowd-sourcing can resolve label conflicts to yield improved performance. While their results re-affirm the value of multi-modal modeling, a limitation is that they mainly focus on the crowd-sourcing mechanism to resolve label conflicts and do not discuss models for modeling multi-modal content. Finally, with the rise of deep-learning, several works have explored the use of distributed representations (word and paragraph embeddings) and large-scale pre-trained models for Tweet topic classification (Li et al., 2016a,b,c,d; Ive et al., 2018; Kang et al., 2019; Gonzalez et al., 2021).

When viewed in sum, one notes at-least one of the following limitations in all of the above works: (a) focus on a very small number of topics $(5 - 20)$ (b) do not support multiple topic labels per Tweet (c) do not consider or discuss how to model content beyond the raw Tweet text (d) do not capture label constraints. A sole exception to many of the above limitations is the work of Yang et al. (2014) which attempts to perform large-scale Tweet topic classification focusing on 300 topic labels in a real-time setting. They only use weakly-labeled data for training a logistic regression model using hashed n-gram based features derived only from the Tweet text. They do not incorporate additional features in their model due to limited training data and very high-dimensional sparse features. Instead in practice, they adopt an integrative approach where they weigh the predictions of the ngram-based model post-hoc using user-topic affinity scores inferred from a closed-loop mechanism.

Here, we revisit their large-scale setting after a decade. Armed with larger and cleaner data, improved content modeling, and the availability of specialized hardware to train large models, we address the challenges faced by (Yang et al., 2014) and propose a vastly improved model for large-scale Tweet topic classification that uses large-scale pre-training and models Tweets holistically.[1]

3 Data

Similar to Yang et al. (2014), we consider a set of 300 topics which are popular and frequently discussed on Twitter.[2] However, while Yang et al. (2014) construct their training data by only using weak labels obtained from a rule-based system using keyword matches, we construct our training data using both (a) high precision human-labeled annotations and (b) weakly-labeled data from a rule-based system using keyword matches.[3]

3.1 Human Annotated Labeled Data (HCOMP Dataset)

Naively sampling a random set of Tweets and asking human annotators to assign potential topics from a list of 300 topics is very inefficient because

[1] In this work, we restrict ourselves to classifying only English Tweets.

[2] See the Appendix for the full list of topics considered.

[3] See the Appendix for a brief description of this keyword-match system for yielding weak labels.

(a) a random sample would yield very few topical Tweets and (b) it imposes a prohibitively large cognitive load on annotators. We mitigate the above limitations by closely following the procedure outlined by (Yang et al., 2014) which first samples Tweets based on topic priors to obtain Tweets that are weakly relevant to a topic, and then seeks label confirmation from trained human annotators. Specifically, we consider Tweets originating from users that are known to tweet mostly about a given topic (for example: Tweets authored by CNN are almost certainly about the "News" topic). We collect 100K such Tweets with at-least 200 Tweets per topic. We then sought label confirmation from trained human annotators with each Tweet-topic pair being independently rated by 3 annotators and use a majority vote to determine the final labels (see Appendix for details). We call this dataset the **HCOMP** dataset. We create training, validation, and test splits of this dataset disjoint at both the Tweet and the user level to ensure an unbiased evaluation.[4]

3.2 Weakly Labeled Data (WLD Dataset)

In addition to the human labeled training data collected above, we also construct a large-scale dataset of weakly labeled Tweets (**WLD** dataset) for task-specific pre-training (see Section 4). Specifically, we use the rule-based system to obtain a random sample of 250 million weakly labeled Tweets.

3.3 Chatter Data (CHT Dataset)

As noted by Yang et al. (2014), our training data should also include Tweets that are non-topical so that our learned model does not incorrectly assign topics to what they term "Twitter chatter" – Tweets that are largely about daily status updates, greetings and clearly non-topical content. Therefore we follow (Yang et al., 2014) and construct a dataset of weakly labeled non-topical Tweets by sampling Tweets that trigger none of the topical rules in the rule-based system. We verify that a small random sample ($N = 150$) of those Tweets (denoted by **CHT-test**) are indeed non-topical through independent human annotators which we set-aside for model evaluation. We use the remaining portion ($N = 100000$) as additional training data, once again ensuring the train and test splits are author and Tweet disjoint.

[4]We do this because as we will see later, we use author level features in our model.

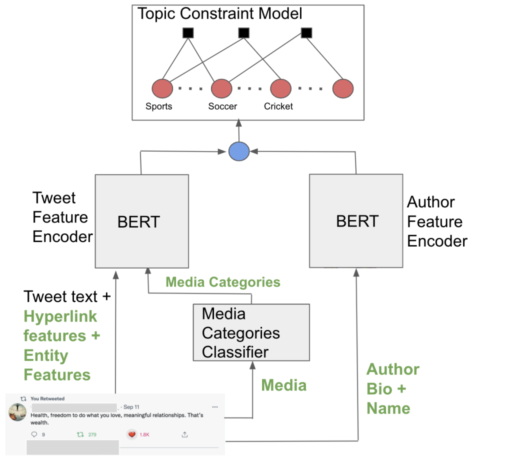

Figure 1: **Overview of our CTM model for large-scale topic classification of Tweets**. Our model consists of 3 components: (a) a Tweet feature encoder encoding Tweet features (b) an Author feature encoder encoding author features thus capturing author-topic affinity and (c) a constraint model that encourages the topic scores to respect prior constraints.

4 Models and Methods

Problem Formulation. We formulate our problem as an instance of standard multi-label classification. Formally, let $\mathcal{S}$ denote the given set of topics. Given $\boldsymbol{X}$, a set of Tweet features and a set of topics $\boldsymbol{L} \in 2^{\mathcal{S}}$, we would like to model $\Pr(\boldsymbol{L}|\boldsymbol{X})$. We encode the topic labels $\boldsymbol{L}$ as a binary vector $\boldsymbol{Y}$ of length $|\mathcal{S}|$ where each $\boldsymbol{Y_i} \in \{0, 1\}$ indicates whether the given Tweet belongs to topic i or not.

We consider a simple approach to multi-label classification – namely model $\Pr(\boldsymbol{Y_i}|\boldsymbol{X})$ for each i.[5] While it is possible to operationalize this by completely independent model artefacts $\Pr(\boldsymbol{Y_i}|\boldsymbol{X})$ each parameterized by $\boldsymbol{\Theta_i}$, this is not a scalable approach when we have 300 topics. Instead, we propose a simple neural architecture parameterized by $\boldsymbol{\Theta}$ that outputs a vector $\hat{\boldsymbol{Y}}$ of length $|\mathcal{S}|$ where each $\hat{\boldsymbol{Y}}_{\boldsymbol{i}} \in [0, 1]$ denotes the probability of belonging to topic i.

Model Overview. Here, we provide an overview of CTM and outline specific details in future sections (also depicted in Figure 1). CTM consists of three main components:

- **Tweet Feature Encoder**: The Tweet feature encoder encodes features of the Tweet holistically. Specifically, it encodes the Tweet text, hyperlink features, named entity mention features, as well as features of associated media.

[5]Other approaches to multi-label classification like classifier-chains and label partitioning approaches are not suitable in our setting since they can be very compute and memory-intensive especially when the the label space is large.

This encoder outputs a vector of topic logits (one for each topic) based on these input features which we denote by $\hat{\boldsymbol{Y}}^t$.

- **Author Feature Encoder**: The author feature encoder encodes author features like the author name and biography which may be indicative of the author's affinity to tweet about certain topics. This encoder outputs a vector of topic logits (one for each topic) based on these input features which we denote by $\hat{\boldsymbol{Y}}^a$. $\hat{\boldsymbol{Y}}^a$ is combined with $\hat{\boldsymbol{Y}}^t$ via a element-wise addition to yield the combined topic logits – $\hat{\boldsymbol{Y}}^c$ which can be converted to probability scores using a sigmoid transformation.
- **Topic Constraint Model**: The topic constraint model encourages the predictions to reflect prior known consistencies and constraints among the topic labels. For example, we would like to encode the notion that Tweets about "Soccer" are almost certainly about "Sports" and very unlikely to also be about "Basketball". Consequently, this component encodes such pre-specified label constraints in the output space via a factor-graph. Performing inference on the factor-graph recalibrates the raw probabilities of the model given by $\hat{\boldsymbol{Y}}^c$ to better reflect the output label constraints yielding the final predicted probabilities for each topic $\hat{\boldsymbol{Y}}^f$.

4.1 Tweet Feature Encoder

The Tweet feature encoder is a standard BERT encoder with a linear classification head where all layers are trainable. Each individual Tweet feature is modeled as follows:

- **Tweet Text**: We simply pass the Tweet text as an input string to BERT after standard preprocessing (case-folding, stripping hyperlinks and user mentions).
- **Hyperlink Features**: For each hyperlink in the Tweet text, we obtain the raw HTML content of the web-page being referenced, and extract the web-page title and the first 100 characters of the web-page description to capture any topical cues. These features are simply concatenated with the Tweet text using a pre-defined separator token.
- **Media**: In order to incorporate topical cues from any attached media (images, gifs, and videos), we obtain media annotations for the given media. The obtained media annotations reflect broad categories that summarize the content of the media. For example, an image with several people playing outside may have the media annotations {MULTIPLE PEOPLE, RUNNING}. To encode such cues captured by the media, we simply concatenate all of these media annotations to the current input string using a pre-defined token as a delimiter. The media annotations themselves are predicted by a media-annotations classifier that learns to assign each media to zero or more categories from a set of pre-defined categories. [6]
- **Entity Features**: Recognizing that mentions of named entities provide strong topical cues (especially when text is short), we extract mentions of named entities in the Tweet text using an off-the-shelf Twitter NER model (Mishra et al., 2021) and link each extracted named entity to their entry in WIKIDATA when available. We use the WIKIDATA descriptions of each linked entity as additional inputs to the Tweet feature encoder. As am example, this enables CTM to infer that Tweets which mention "Steve Waugh" are likely about "Cricket".

Pretraining the Tweet Feature Encoder. Noting that the weights of the standard BERT encoder are not reflective of the domain of Tweets and may represent a poor initialization point during subsequent finetuning, we pretrain the BERT encoder on the task of predicting topics using the **WLD** dataset only using the raw Tweet text as the input feature. It is important to note that the set of Tweets used for pre-training is completely disjoint from the test-set and the training set (used in full model training) both in terms of time-span and Tweets. As we will show empirically, this large-scale pretraining helps our model improve generalization performance and suggests that the full model is able to to better adapt to the domain of Twitter data and the specific task. These observations are also inline with the findings of Gururangan et al. (2020) who also note the effectiveness of task and domain specific pretraining.

4.2 Author Feature Encoder

The author feature encoder is also identical to a standard BERT encoder with a linear classification head with all layers being trainable. We use the

[6]See the Appendix for details on the media categories classifier.

following features of the author (all of which are simply concatenated together as input to BERT): (a) **Author Biography**: We use the self-reported publicly available author-profile description of the author posting the Tweet. (b) **Author Name**: We also use the author's display name. We hypothesize that all of these features may be indicative of the topics that the author likely tweets about. For example, an author name containing the string "FashionNews" strongly suggests that Tweets made by that author will likely be about Fashion.

Pretraining the Author Feature Encoder. Similar to pre-training the Tweet feature encoder, we also perform pre-training of the author feature encoder using only the author-biography feature of a large set of authors. More specifically, we consider a sample of 100K authors (disjoint from the test set) and obtain their background Tweets during a historical time-period so that there is no overlap with the test sets. Given these background Tweets of a user, we use a keyword-based approach to obtain weak topic labels for each of those Tweets. This allows us to compute a histogram of topics associated with the user and reflects the affinity of the user to Tweet about specific topics. By thresholding these topic affinity scores we estimate the most likely topics associated with a given author. The author-encoder is then pre-trained to predict the topics associated with the author using only their biography.

4.3 Topic Constraint Model

The topic constraint model encodes output label constraints in the topic prediction and captures correlations among topics. We encode such dependencies (hard or soft) via the framework of factor graphs. In particular, given a vector of topic predictions (probabilities) $\hat{\boldsymbol{Y}}^c$, for each topic T_i, we associate a discrete binary random variable with that topic v_i, and a corresponding unary factor with potential function f_i such that $f_i(0) = 1.0 - \hat{\boldsymbol{Y}}_{\boldsymbol{i}}^{c}$ and $f_i(1) = \hat{\boldsymbol{Y}}_{\boldsymbol{i}}^{c}$. For every constraint between a pair of topics (i, j), we construct a binary factor with potential function $\phi_{i,j}(v_i, v_j)$. This potential function encodes the compatibility between prediction scores for topic i and topic j. Domain experts can craft their own potential functions to reflect positive or negative compatibility between topic pairs or alternatively even learn these from correlation data. In this work, since we have strong intuitions on the constraints we hand-craft the appropriate potential functions. We consider two types of constraints:

- **Broader Topic Inclusion**: If a Tweet is about a specific topic c, then it is very likely that the Tweet is also about topic p where p subsumes topic c. Other cases are a "don't-care". This encodes the notion that some topics tend to be active at the same time. For example, if a Tweet is about "Basketball", it is almost certainly about "Sports". It is important to the note that this relation can be asymmetric. We use the following potential matrix[7] for encoding this type of constraint:

p \ c	**0**	**1**
0	0.5	0.0
1	0.5	10.0

- **Topic Pair Exclusion**: At-most one among topic x and y can be active at any time. This encodes mutually exclusive topics. For example, it is very unlikely to have a Tweet which is about both Cricket and Basketball. We use the following potential matrix for encoding this type of constraint:

x \ y	**0**	**1**
0	0.5	0.5
1	0.5	0

Having constructed a factor graph encoding the specified output constraints, we then use belief propagation on the factor graph to obtain updated marginal probabilities for each of our variables (topics) which reflect the encoded output constraints. Specifically, messages are alternately passed between variable nodes and factor nodes (until convergence is achieved or a finite number of iterations is completed). A message is simply a vector μ where the individual components denote the probability of the random variable taking a specific value $x \in \{0, 1\}$. The message from a variable v to neighboring factor f on taking a specific value x is given by the following equation:

$$\mu_{v \longrightarrow f}(x) \propto \prod_{g \in \mathcal{N}(v) \setminus f} \mu_{g \longrightarrow v}(x) \qquad (1)$$

, where g belongs to the set of factor nodes connected to v excluding f. Similarly, the message

[7]The potential matrices are not necessarily unique and other equivalent matrices may exist.

from a factor node f to the variable v on the variable taking a specific value x is given by the following:

$$\mu_{f \longrightarrow v}(x) \propto \sum_{\mathbf{x}:\mathbf{x}_v = x} \phi(\mathbf{x}) \prod_{u \in \mathcal{N}(f) \backslash v} \mu_{u \longrightarrow f}(\mathbf{x}_u) \tag{2}$$

, where u belongs to the set of variable nodes connected to f excluding v.

Finally, after convergence (or a finite number of iterations), the updated marginal probability of variable v taking on a value x is given by $\Pr(v = x) \propto \prod_{g \in \mathcal{N}(v)} \mu_{g \longrightarrow v}(x)$. This yields $\hat{\boldsymbol{Y}}^f$, the final output predicted probabilities. In our experiments, we impose the above constraint types on specific topics falling under (and including) the broad topics of Sports, Music, Animation, Science, Animals, Anime & Manga.

4.4 Parameter Estimation

During training, we minimize binary cross-entropy loss (where the loss for an instance is computed for each topic independently and summed over all topics). Since for a given instance, any topic which not explicitly marked as a positive is treated implicitly as a negative there is high class imbalance. Therefore, we employ positive class weighting where the class weight for topic c is computed as fraction of negative examples to positive examples of topic c in the training data.

5 Experiments

Here, we present results on evaluating our models both quantitatively and qualitatively.

5.1 Quantitative Evaluation

Baselines and Evaluation Setup. We consider two baselines: (a) A bag-of-words logistic regression (LR) model – our best-effort attempt to reproduce the decade old setup of Yang et al. (2014) which uses only the tweet text (b) Replace logistic regression with a standard BERT model using the Tweet text. We train all models on the training data set, and evaluate them on the two held-out test sets:

- **HCOMP Test Set**: We evaluate model performance on the held out test split from the **HCOMP** dataset. We report the median average precision score over all topics. We consider the average precision score, since unlike the F1 score, it summarizes model performance over all operating thresholds.
- **CHT Test Set**: In order to measure the ability of our models to effectively reject assigning topics to "non-topical" Tweets (chatter), we evaluate our models on the held-out chatter test set. Here, we report the number of predictions made by the model over a given probability threshold (we use 0.9 but our results hold for other thresholds as well). The lower the number of model predictions above a reasonable operating threshold, the better the model is at ensuring non-topical Tweets are not assigned topics.

We also perform a systematic feature ablation study of our proposed CTM model to quantify the effect of feature sets considered. Table 1 shows the results of our evaluation. Note that our full CTM significantly outperforms the logistic regression and BERT baselines (**Median APS**: $\mathbf{67.0}$ vs 54.8) and yields a relative improvement of $\mathbf{20}\%$ thus underscoring the effectiveness of our approach. We also make the following observations based on our ablation experiments:

- **Including Tweets from the CHT dataset improves performance at detecting chatter**. This claim is supported by noting the performance on the **CHT** test-set where the BERT model trained on chatter Tweets shows better performance than the same model trained without ($\mathbf{135}$ vs $\mathbf{254}$).
- **Media features have a focused impact.** Note that adding media annotations overall does not affect the median average precision score significantly (see row **CTM-A**: $\mathbf{54.4}$ vs $\mathbf{54.8}$). However, observe that many tweets in the evaluation may not contain media annotations. If we restrict our evaluation to only the tweets containing media, we observe a significant improvement by modeling media. The corresponding average precision scores are $\mathbf{71.0}$ vs $\mathbf{58.4}$. This observation suggests that media features have a targeted and focused impact and are expected to help topics that tend to be media rich. To further characterize this focused impact of media features, we ranked the individual topics (with at-least 100 positive test examples) based on their performance improvement due to media annotations. We observed that that media features significantly boost the performance of Automotive, US national news, Anime, and Movies which indeed tend to be media rich.

	Setting	Median APS ↑	CHT ↓
LR(baseline) (Yang et al., 2014)	Tweet text (trained on only HCOMP)	33.0	108
BERT(baseline)	Tweet text (trained on only HCOMP)	54.5	254
BERT (baseline)	Tweet text (trained on HCOMP + CHT)	54.8	135
CTM-A	Tweet text + media annotation (trained on HCOMP + CHT)	54.4	121
CTM-B	CTM-A + pretraining	56.7	107
CTM-C	CTM-B + Hyperlink features	57.2	101
CTM-D	CTM-C + User features	63.3	75
CTM-E	CTM-D + Entity Linking features	66.5	80
CTM-F (Full model)	CTM-E + Constraint model	67.0	90

Table 1: **Performance of CTM on the test sets.** The median APS is the median average precision on the **HCOMP** test set (*higher is better*) where as **CHT** column shows the number of model predictions exceeding a probability score of 0.9 on the **CHT** test set (*lower is better*). Note that CTM significantly outperforms baseline models and demonstrates the effectiveness of modeling content beyond the immediate Tweet text.

Topic	APS (w/o constraint model)	APS (with constraint model)
Animation	0.64	0.71
Animals	0.88	0.91
Anime & manga	0.66	0.84
Music	0.41	0.70
Sports	0.69	0.89
Science	0.44	0.63

Table 2: **Performance improvements due to the constraint model.** The constraint model yields significant improvements on broader topics (as large as 20 points). Performance on narrower topics do not change significantly.

Tweet Content	Predicted Label	Helpful feature
In times of trouble, regression models come to me, speaking words of wisdom	Data Science	Tweet text
Power hitter joins #yellowstorm `att:Attached media of cricket bat and gloves`	Cricket	Media Annotations
Cameras in USC vs UT stopped working, so it is a podcast now	American Football	Author Bio
Revealed: Australia's stars set to be pulled from IPL `URL to fox.sports domain`	Cricket	Hyperlink
cody ko and noel miller are just ...	Digital creators	Entity features

Table 3: A few examples of correct model predictions that also illustrate the benefit of different feature sets. Tweets are paraphrased to protect user privacy.

Tweet Content	Predicted Label	Error Reason
In life, you have not seen your best days, you have not run your best race ...	Running	Metaphor
Cheerleading the mob is not going to save ...	Cheerleading	Metaphor
I am going to have very large drink tonight not sure if whisky or cyanide	Food	Sarcasm or Irony
I need my **** ate	Food	NSFW sense
This is a thread 1/5...	No topic	Conversation thread
On this day of Buddha Purnima..	Yoga	Close topic

Table 4: A few challenging cases for our model. Tweets are paraphrased to protect user privacy.

- **Large-scale pretraining of feature encoders boosts overall performance.** Observe that pre-training the encoders on domain (and task) specific data is very effective (see row **CTM-B:Median APS – 56.7 vs 54.4**). This observation reaffirms findings of (Gururangan et al., 2020) advocating the effectiveness of domain and task specific pre-training.
- **Hyperlink features have a focused impact.** With regards to hyperlink features, we observe that hyperlink features are generally useful but have a negligible overall impact (see row **CTM-C:Median APS – 57.2 vs 56.7**). Similar to our analysis of performance of media features, if we restrict our evaluation to only those instances with. hyperlinks we indeed observe a significant performance gain by incorporating hyperlink features. Specifically the corresponding scores are **92.67** vs **83.4**. A closer analysis reveals that hyperlink features most improve the performance on Travel, Movies, Gaming, and US national news – topics that tend to be hyperlink heavy.
- **Author features significantly boost overall performance.** Author features yield the most benefit overall (see row **CTM-D:Median**

APS – 63.3 vs 57.2). This observation reaffirms the importance of user-level modeling in NLP tasks (especially on social media) and supports observations made by Lynn et al. (2019) who show that author features can improve performance on a variety of tasks like stance detection, and sarcasm detection.

- **Entity features also significantly boost overall performance.** Similar to author features, the entity features also significantly improve overall performance (see row **CTM-E:Median APS – 66.5 vs 63.3**). Drilling down, we noted that entity linking features most improve the performance on Rap, American football, K-pop, Entertainment News, and Cricket – all topics whose Tweets are likely to mention specific sport players, movie stars, and musicians that are suggestive of the topic.
- **The constraint model significantly boosts the performance of the relevant topics.** Finally, we note that including the constraint model only very slightly improves the median average precision score (67.0 vs 66.5). This is to be expected since the constraint model only affects topics for which constraints were included. Therefore, it is illustrative to look at the performance on the topics for which constraints were imposed as shown in Table 2. Note that while we observe a slight degradation in the performance on **CHT** dataset by using the constraint model (90 vs 80)[8] we see a very significant increase in the average precision score of all the broad topics for which constraints are introduced (by as much as 20 points) because the model reduces constraint violations – especially violations of the broader topic inclusion constraint. For example, the model correctly infers that Tweets about specific sports like "Basketball" are also very likely about "Sports" thus significantly boosting performance on the Sports topic (0.89 vs 0.69). Similar improvements are also seen for other topics like Music etc.

5.2 Qualitative Evaluation

In addition to evaluating our CTM quantitatively, we also inspected the model predictions qualitatively to identify instances which (a) reveal the benefit of various features resulting in a correct prediction and (b) highlight challenging cases where the model still struggles. Table 3 shows a few sample Tweets where the model predictions are correct and illustrates the benefit of modeling Tweet content holistically. In "Power hitter joins #yellowstorm", only the attached media (which displays a cricket apparel) is indicative of the topic. Similarly, our model correctly predicts that "Revealed: Australia's stars set to be pulled from IPL" is about "Cricket" by leveraging topical cues extracted from the linked website's description and title. Finally, CTM correctly infers that the Tweet referencing "Cody Ko and Noel Miller" is about "Digital Creators" by leveraging named entity cues.

While CTM undoubtedly advances the state of the art, we observe a few systematic failure modes shown in Table 4. We note five challenging areas that suggest future directions for improved modeling: (a) **Metaphorical Usage**: Our model is unable to pick up on metaphorical usage of topical words like "running" or "cheer-leading". (b) **Sarcasm or Irony**: CTM does not pick up on sarcasm and assumes topical content when none is intended. (c) **NSFW Senses**: Our model finds it challenging to distinguish between NSFW senses of certain phrases (and words) and their general topical meanings. (d) **Threads**: Our model is unable to infer topics for tweet threads because we do not model conversations. (e) **Close Topics**: We also note a few cases where the model is unable to distinguish between close and related topics sharing topical keywords (for eg. Buddhism and Yoga).

6 Conclusion

We revisited the problem of large scale Tweet topic classification posed by (Yang et al., 2014) and proposed a model for classifying Tweets into a large set of 300 topics with improved performance. In addition to tackling a significantly larger topic set than prior work, our model takes a holistic approach to modeling Tweets. We model not only the immediate Tweet text, but also associated media, hyperlinks, author context, entity mentions. Our model can also incorporate domain knowledge expressed in the form of topic constraints in a principled manner. Our holistic approach to large-scale Tweet topic modeling thus sets the stage for improved Tweet annotation models which can significantly improve downstream recommendation systems and search engines in social media platforms to enhance user experience.

[8] This slight degradation on CHT is due to error propagation of high confidence false positives which occurs to respect the constraints.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*.

Marko Balabanovic and Yoav Shoham. 1995. Learning information retrieval agents: Experiments with automated web browsing. In *On-line Working Notes of the AAAI Spring Symposium Series on Information Gathering from Distributed, Heterogeneous Environments*, pages 13–18.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Harold Borko and Myrna Bernick. 1963. Automatic document classification. *Journal of the ACM (JACM)*, 10(2):151–162.

Shuo Chang, Peng Dai, Jilin Chen, and Ed H Chi. 2015. Got many labels? deriving topic labels from multiple sources for social media posts using crowdsourcing and ensemble learning. In *Proceedings of the 24th International Conference on World Wide Web*, pages 397–406.

Héctor Cordobés, Antonio Fernández Anta, Luis F Chiroque, Fernando Pérez, Teófilo Redondo, and Agustín Santos. 2014. Graph-based techniques for topic classification of tweets in spanish. *International Journal of Interactive Multimedia and Artificial Intelligence*.

Yegin Genc, Yasuaki Sakamoto, and Jeffrey V Nickerson. 2011. Discovering context: classifying tweets through a semantic transform based on wikipedia. In *International conference on foundations of augmented cognition*, pages 484–492. Springer.

Jose Angel Gonzalez, Lluís-F Hurtado, and Ferran Pla. 2021. Twilbert: Pre-trained deep bidirectional transformers for spanish twitter. *Neurocomputing*, 426:58–69.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Julia Ive, George Gkotsis, Rina Dutta, Robert Stewart, and Sumithra Velupillai. 2018. Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 69–77.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.

Jaeyong Kang, HongSeok Choi, and Hyunju Lee. 2019. Deep recurrent convolutional networks for inferring user interests from social media. *Journal of Intelligent Information Systems*, 52(1):191–209.

Saurabh Kataria and Arvind Agarwal. 2015. Supervised topic models for microblog classification. In *2015 IEEE International Conference on Data Mining*, pages 793–798. IEEE.

Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. 2011. Twitter trending topic classification. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 251–258. IEEE.

Quanzhi Li, Sameena Shah, Mohammad Ghassemi, Rui Fang, Armineh Nourbakhsh, and Xiaomo Liu. 2016a. Using paraphrases to improve tweet classification: Comparing wordnet and word embedding approaches. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 4014–4016. IEEE.

Quanzhi Li, Sameena Shah, Xiaomo Liu, Armineh Nourbakhsh, and Rui Fang. 2016b. Tweet topic classification using distributed language representations. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 81–88. IEEE.

Quanzhi Li, Sameena Shah, Xiaomo Liu, Armineh Nourbakhsh, and Rui Fang. 2016c. Tweetsift: Tweet topic classification based on entity knowledge base and topic enhanced word embedding. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 2429–2432.

Quanzhi Li, Sameena Shah, Armineh Nourbakhsh, Xiaomo Liu, and Rui Fang. 2016d. Hashtag recommendation based on topic enhanced embedding, tweet entity data and learning to rank. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 2085–2088.

Veronica Lynn, Salvatore Giorgi, Niranjan Balasubramanian, and H Andrew Schwartz. 2019. Tweet classification without the tweet: An empirical examination of user versus document attributes. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 18–28.

Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing demographic bias in named entity recognition. *arXiv preprint arXiv:2008.03415*.

Shubhanshu Mishra et al. 2021. Improved multilingual language model pretraining for social media text via translation pair prediction. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 381–388, Online. Association for Computational Linguistics.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

P Selvaperumal and A Suruliandi. 2014. A short message classification algorithm for tweet classification. In *2014 International Conference on Recent Trends in Information Technology*, pages 1–3. IEEE.

Anna Stavrianou, Caroline Brun, Tomi Silander, and Claude Roux. 2014. Nlp-based feature extraction for automated tweet classification. *Interactions between Data Mining and Natural Language Processing*, 145.

Ke Tao, Fabian Abel, Claudia Hauff, and Geert-Jan Houben. 2012. What makes a tweet relevant for a topic? In *# MSM*, pages 49–56. Citeseer.

Kimitaka Tsutsumi, Kazutaka Shimada, and Tsutomu Endo. 2007. Movie review classification based on a multiple classifier. In *Proceedings of the 21st pacific Asia conference on language, information and computation*, pages 481–488.

Shuang-Hong Yang, Alek Kolcz, Andy Schlaikjer, and Pankaj Gupta. 2014. Large-scale high-precision topic modeling on twitter. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1907–1916.

A Appendix

A.1 Details Regarding Off the Shelf Components Used in CTM

A.1.1 Media Annotations Classifier

The media annotations classifier takes as input an image and classifies the image into one or more of 45 media categories listed in Table 5. The classifier is essentially a standard MOBILENET V2 model (Sandler et al., 2018) further fine-tuned on a human-labeled curated dataset of 100K images from Twitter. The operating threshold of the media classifier is set to achieve a precision of about 90% on each topic.[9]

A.1.2 Twitter Named Entity Recognizer

The Twitter NER model is a standard bi-directional LSTM with a CRF layer and detects mentions of persons, places, organizations, and products in a Tweet. The model has been trained on 100K human annotated labeled tweets (Mishra et al., 2020) and has a precision of 85% with a recall of 70% on a held-out test set. We link the extracted mention to a potential WikiData candidate as follows: (a) we first construct a set of potential WikiData entity candidates - the set of all entities whose label or alias has a match with the extracted mention (b) link the mention to the top entity candidate obtained by sorting the candidate set in descending order of page view count as the primary key breaking ties using page rank as the secondary key. We use this approach as an expedient choice noting that more sophisticated entity linking approaches can be used.

A.1.3 Rule Based System for Generating Weakly Labeled Examples.

We employ a rule-based system consisting of tens of thousands of rules based on key-words to generate weakly labeled examples. All rules are manually curated and added by domain experts and data specialists.

A.2 Hyper-parameter Tuning

We explored several hyper-parameter settings for the baseline models namely Logistic Regression and BERT to make baseline comparisons strong and compare CTM against only the best performing baseline settings. In particular, we explored training for different epochs $(1-10)$ for the BERT baseline. For the logistic regression baseline, we also tried various settings for the maximum number of iterations of the optimizer $(100-1000)$ as well as various values for the strength of the L2 regularizer ($C = [0, 1, 10, 100]$).

For our proposed model CTM, we did not do any specific hyper-parameter tuning and just trained all models for 5 epochs using 1 A100 GPU.

A.3 Details on the Human Labeled Annotation Task

In this section, we briefly describe the human annotation task used for obtaining topic label confirmation used in the construction of the **HCOMP** dataset. Each annotator is shown a Tweet, topic pair and asked to judge whether the topic is relevant to the Tweet or not. The instructions are:

> **Task**: In this task, you will be shown a tweet and a topic and asked whether the tweet is 'relevant' for a topic.
>
> **Topics**:You will be asked to determine if a tweet is relevant for a given topic. A "Topic" is a potential subject of conversation that can be identified with a commonly held definition, where mass interest in the subject is not likely to be temporary, e.g. 'Comedy' or 'Knitting' is a topic as it is non-subjective and has a commonly held definition. Purely social tweets like "are you doing okay?" or personal remarks like "I'm having a bad day" are not topical. A Tweet can be popular without being topical.
>
> **Question**: The primary question you will be asked is "Is this tweet about a topic?", the possible responses are: Yes - This tweet is primarily about this topic. Somewhat - This tweet is related to this topic, but it is not a primary topic of this tweet. No - This tweet is unrelated to this topic. Unsure - I don't understand this tweet.
>
> **Guidelines**: You will first want to make sure you understand the presented topic. If you are unfamiliar with the topic presented in this question, please click on the topic which will take you to a Google search result page. Feel free to click on a few links (news articles or a Wikipedia page) to familiarize yourself with the topic. When elements of the tweet can I use to make a judgment? It can sometimes be challenging to tell what a tweet is about from tweet text alone. In order to determine what the tweet is about you may need to do the following: Look at replies of a tweet, which might provide additional context by clicking on the tweet. (NOTE: If you can understand the tweet by relying just on the body or author of the tweet, it is fine to not designate replies as being used to make a judgment.) Google phrases in the tweet text if you are unfamiliar with a mentioned entity or phrase that will help you understand the tweet. Look at the image, video, or click on any link (including a hashtag) associated with the Tweet, since it may be commenting on this media. If the media is primarily about the topic, the tweet is as well. Look at the tweet author's name, profile, public timeline, or linked website if it helps disambiguate tweet content. (NOTE: Please don't use the author alone in making determination, without some other element of the tweet.)

Each HIT is judged by 3 independent highly reliable annotators.

[9]For videos, and GIF's each frame is analyzed by the model with the prediction scores being aggregated using the max operator.

A.4 Data Statement

Here, we outline other aspects of our data as per recommendations outlined in (Bender and Friedman, 2018).

SUMMARY – We collect a set of tweet, topic pairs focusing on only English Tweets which we use for predictive modeling and evaluation.

CURATION RATIONALE – The rationale for the setup used in data collection was primarily driven by our task (large scale topic classification) and the need for data to a build a predictive model. The size of the data collected was thus influenced by task, available budget, and time available.

LANGUAGE VARIETY - The tweets were restricted to English only and are from the time range between September 2020 and May 2021. More fine-grained information is not available.

SPEAKER DEMOGRAPHIC – We do not have any demographic information of the users in this data. One would expect the demographic information to be similar to the demographics of Twitter users around the time of data collection.

ANNOTATOR DEMOGRAPHIC – Human Annotators are primarily native English speakers. No other information is available.

TEXT CHARACTERISTICS – Tweets are short informal and have at-most 280 characters. Tweets are generally meant to be engaged with by other Twitter users.

App Screenshots	Entertainment Events	Pets
Arts and Crafts	Food	Piercing
Auto Racing	American Football	Running
Automotive	Gambling	Single Person
Baseball	Gaming	Skateboarding
Basketball	Golf	Skiing
Beauty, Style and Fashion	Hockey	Smoking
Boxing	Home and Garden	Pharmaceuticals and Healthcare
Captioned Images	Infographics, Text and Logos	Snowboarding
Comics, Animation and Anime	Martial Arts	Soccer
Cricket	Multiple People	Swimming
Crowds and Protests	Nature and Wildlife	Tennis
Currency	Weapons	Travel
Cycling	Other	TV Broadcasts
Drinks	Performance Arts	Weather and Natural Disasters

Table 5: List of 45 media categories that make up the label space of the media classifier.

2D animation	Country music	Horses	Rock climbing
3D animation	Cricket	Hotels	Rodeo
Accounting	Cruise travel	Houston	Roleplaying games
Action and adventure films	Cult classics	Independent films	Romance books
Adventure travel	Curling	Indie rock	Rowing
Advertising	Cybersecurity	Information security	Rugby
Agriculture	Cycling	Interior design	Running
Air travel	Dance	Internet of things	Sailing
Alternative rock	Darts	Investing	Saxophone
American football	Data science	J-pop	Sci-fi and fantasy
Animals	Databases	Jazz	Sci-fi and fantasy films
Animated films	Dating	Jewelry	Science
Animation	Digital creators	Job searching and networking	Science news
Animation software	Documentary films	Judo	Screenwriting
Anime	Dogs	K-hip hop	Sculpting
Anime & manga	Drama films	K-pop	Sharks
Antiques	Drawing and illustration	Kaiju	Shoes
Archaeology	Drums	Knitting	Shopping
Architecture	EDM	Lacrosse	Skateboarding
Art	Economics	Language learning	Skiing
Artificial intelligence	Education	Latin pop	Skin care
Arts& culture	Electronic music	MMA	Small business
Arts & culture news	Entertainment	Makeup	Sneakers
Arts and crafts	Entertainment news	Marine life	Snooker
Astrology	Environmentalism	Marketing	Soap operas
Astronauts	Esports	Martial arts	Soccer
Athletic apparel	Europe travel	Mathematics	Soccer stats
Augmented reality	Everyday style	Men’s boxing	Soccer transfers
Australian rules football	Experimental music	Men’s golf	Soft rock
Auto racing	Famous quotes	Men’s style	Softball
Automotive	Fantasy baseball	Motorcycle racing	Space
Aviation	Fantasy basketball	Motorcycles	Sporting goods
Backpacking	Fantasy football	Movie news	Sports
Badminton	Fantasy sports	Movies	Sports news
Ballet	Fashion	Movies & TV	Sports stats
Baseball	Fashion and beauty	Museums	Startups
Basketball	Fashion business	Music	Storyboarding
Beauty	Fashion magazines	Music festivals	Street art
Biographies and memoirs	Fashion models	Music industry	Streetwear
Biology	Fast food	Music news	Supernatural
Biotech and biomedical	Fiction	Music production	Surfing
Birdwatching	Fighting games	Musicals	Swimming
Black Lives Matter	Figure skating	Mystery and crime books	Table tennis
Blues music	Financial services	National parks	Tabletop gaming
Board games	Fintech	Nature	Tabletop role-playing games
Bollywood dance	Fishing	Nature photography	Tattoos
Bollywood films	Fitness	Netball	Tech news
Bollywood music	Folk music	Nonprofits	Technology
Bollywood news	Food	Olympics	Television
Books	Food inspiration	Online education	Tennis
Bowling	Futurology	Open source	Theater
Boxing	Game development	Opera	Theme parks
Brazilian funk	Gaming	Organic	Thriller films
Business & finance	Gaming news	Organic foods	Track & field
Business media	Gardening	Outdoor apparel	Trading card games
Business news	Genealogy	Outdoors	Traditional games
Business personalities	Geography	Painting	Travel
C-pop	Geology	Parenting	Travel guides
Careers	Golf	Pets	Travel news
Cartoons	Graduate school	Philosophy	Triathlon
Cats	Grammy Awards	Photography	US national news
Cheerleading	Graphic design	Physics	Veganism
Chemistry	Guitar	Podcasts & radio	Vegetarianism
Chess	Gymnastics	Poker	Venture capital
Classic rock	Hair care	Pop	Video games
Classical music	Halloween films	Pop Punk	Visual arts
Cloud computing	Handbags	Pop rock	Volleyball
Cloud platforms	Hard rock	Progressive rock	Watches
College life	Health news	Psychology	Weather
Combat sports	Heavy metal	Punjabi music	Web development
Comedy	Historical fiction	Punk	Weddings
Comedy films	History	R&B and soul	Weight training
Comics	Hockey	Rap	Women’s boxing
Computer programming	Home & family	Reality TV	Women’s golf
Concept Art	Home improvement	Reggae	Women’s gymnastics
Construction	Horoscope	Reggaeton	World news
Cooking	Horror films	Road trips	Wrestling
Cosplay	Horse racing and equestrian	Rock	Yoga

Table 6: List of topics comprising our label space.