Categorizing and Inferring the Relationship between the Text and Image of Twitter Posts

Alakananda Vempala
Bloomberg LP
avempala@bloomberg.net

Daniel Preoţiuc-Pietro
Bloomberg LP
dpreotiucpie@bloomberg.net

Abstract

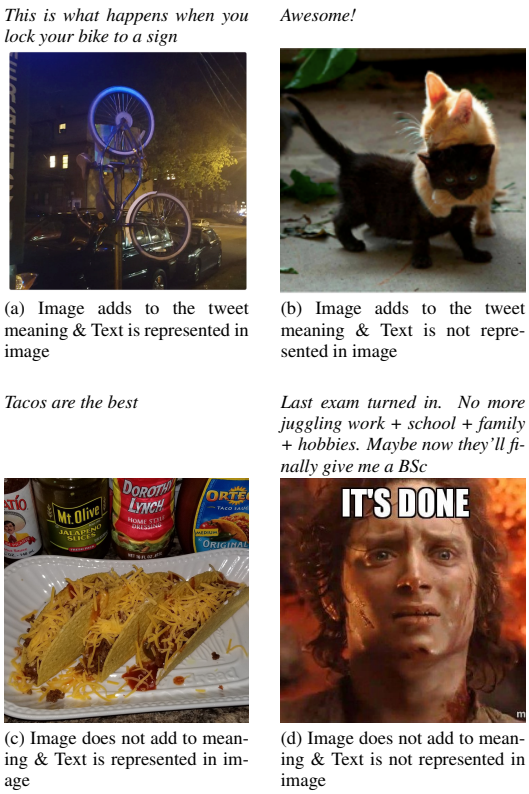
Text in social media posts is frequently accompanied by images in order to provide content, supply context, or to express feelings. This paper studies how the meaning of the entire tweet is composed through the relationship between its textual content and its image. We build and release a data set of image tweets annotated with four classes which express whether the text or the image provides additional information to the other modality. We show that by combining the text and image information, we can build a machine learning approach that accurately distinguishes between the relationship types. Further, we derive insights into how these relationships are materialized through text and image content analysis and how they are impacted by user demographic traits. These methods can be used in several downstream applications including pre-training image tagging models, collecting distantly supervised data for image captioning, and can be directly used in end-user applications to optimize screen estate.

1 Introduction

Social media sites have traditionally been centered around publishing textual content. Recently, posting images on social media has become a very popular way of expressing content and feelings especially due to the wide availability of mobile devices and connectivity. Images are currently present in a significant fraction of tweets and tweets with images get double the engagement of those without (Buffer, 2016). Thus, in addition to text, images have become key components of tweets.

However, little is known about how textual content is related to the images with which they appear. For example, concepts or feelings mentioned in text could be illustrated or strengthened by images, text can point to the content of an image or can just provide commentary on the image content. Formalizing and understanding the relationship between the two modalities – text and images – is useful in several areas:

a) for NLP and computer vision research, where image and text data from tweets are used to developing data sets and methods for image captioning (Mitchell et al., 2012) or object recognition (Mahajan et al., 2018);
b) for social scientists and psychologists trying to understand social media use;
c) in browsers or apps where images that may not contain additional content in addition to the text would be replaced by a placeholder and displayed if the end-user desires to in order to op-



(a) Image adds to the tweet meaning & Text is represented in image

(b) Image adds to the tweet meaning & Text is not represented in image

(c) Image does not add to meaning & Text is represented in image

(d) Image does not add to meaning & Text is not represented in image

Figure 1: Examples of the four types of text-image relationship from this study.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2830–2840
Florence, Italy, July 28 - August 2, 2019.

timize screen space (see Figure 2).

Figure 1 illustrates four different ways in which the text and image of the same tweet can be related:

- Figures 1(a,b) show how the image can add to the semantics of the tweet, by either providing more information than the text (Figure 1a) or by providing the context for understanding the text (Figure 1b);
- In Figures 1(c,d), the image only illustrates what is expressed through text, without providing any additional information. Hence, in both of these cases, the text alone is sufficient to understanding the tweet's key message;
- Figures 1(a,c) show examples of tweets where there is a semantic overlap between the content of the text and image: *bike* and *sign* in Figure 1a and *tacos* in Figure 1c;
- In Figures 1(b,d), the textual content is not represented in the image, with the text being either a comment on the image's content (Figure 1b) or the image illustrating a feeling related to the text's content.

In this paper, we present a comprehensive analysis that focuses on the types of relationships between the text and image in a tweet. Our contributions include:

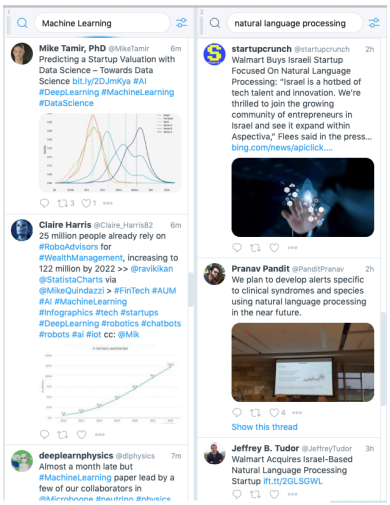
- Defining the types of relationships between the text and the image of a social media post;
- Building a data set of tweets annotated with text - image relationship type;[1]
- Machine learning methods that use both text and image content to predict the relationship between the two modalities;
- An analysis into the author's demographic traits that are related to usage preference of text-image relationship types;
- An analysis of the textual features which characterize each relationship type.

2 Related Work

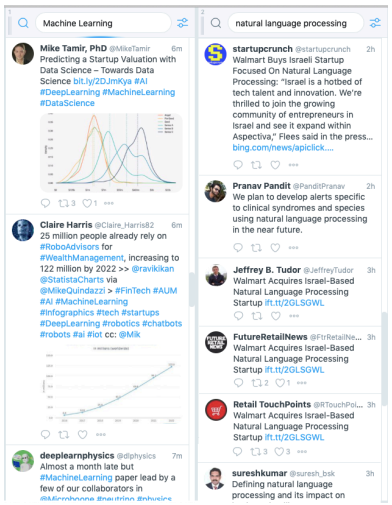
Task. The relationship between a text and its associated image was researched in a few prior studies. For general web pages, Marsh and Domas White (2003) propose a taxonomy of 49 relationship grouped in three major categories based on how similar is the image to the text ranging from little relation to going beyond the text, which forms the basis of one of our relationship dimensions. Martinec and Salway (2005) aim to categorize text-image relationships in scientific articles from two perspectives: the relative importance of one modality compared to the other and the logico-semantic overlap. Alikhani and Stone (2018) argue that understanding multimodal text-image presentation requires studying the coherence relations that organize the content. Even when a single relationship is used, such as captioning, it can be expressed in multiple forms such as telic, atelic or stative (Alikhani and Stone, 2019). Wang et al. (2014) use the intuition that text and images from microposts can be associated or not or depend on one another and use this intuition in a topic model that learns topics and image tags jointly. Jas and Parikh (2015) study the concept of image specificity through how similar to each other are multiple descriptions of that image. However, none of these studies propose any predictive methods for text-image relationship types. Alikhani et al. (2019) annotate and train models on a recipe data set (Yagcioglu et al., 2018) for the relationships between instructional text and images around the following dimensions: temporal, logical and incidental detail. Chen et al. (2013) study text-image relationships using social media data focusing on the distinction between images that are overall visually relevant or non-relevant to the textual content. They build models using the text and image content that predict the relationship type (Chen et al., 2015). We build on this research and define an annotation scheme that focuses on each of the two modalities separately and look at both their semantic overlap and contribution to the meaning of the whole tweet.

Applications. Several applications require to be able to automatically predict the semantic text-image relationship in the data. Models for automatically generating image descriptions (Feng and Lapata, 2010; Ordonez et al., 2011; Mitchell et al., 2012; Vinyals et al., 2015; Lu et al., 2017) or predicting tags (Mahajan et al., 2018) are built using large training data sets of noisy image-text pairs from sources such as tweets. Multimodal named entity disambiguation leverages visual context vectors from social media images to aid named entity disambiguation (Moon et al., 2018). Multimodal topic labeling focuses on generating candidate labels (text or images) for a given topic and ranks them by relevance (Sorodoc et al., 2017). Several resources of images paired

[1]Data set is available at: https://github.com/danielpreotiuc/text-image-relationship/



(a) Full feed with all images displayed



(b) Feed which hides images that do not add content

Figure 2: Example of application using the image task classifier. Automatically collapsing images that do not add content beyond text optimizes screen real estate and allows users to view more tweets in their feed view. The end-user could open hidden images individually.

with descriptive captions are available, which can be used to build similarity metrics and joint semantic spaces for text and images (Young et al., 2014). However, all these assume that the text an image represent similar concepts which, as we show in this paper, is not true in Twitter. Being able to classify this relationship can be useful for all above-mentioned applications.

3 Categorizing Text-Image Relationships

We define the types of semantic relationships that can exist between the content of the text and the image by splitting them into two tasks for simplicity. The first task is centered on the role of the text to the tweet's semantics, while the second focuses on the image's role.

The first task – referred to as the **text task** in the rest of the paper – focuses on identifying if there is semantic overlap between the context of the text and the image.

This task is the defined using the following guidelines:

1. Some or all of the content words in the text are represented in the image (**Text is represented**)
2. None of the content words in the text are represented in the image (**Text is not represented**):
   - None of the content words are represented in the image, or
   - The text is only a comment about the content of the image, or
   - The text expresses a feeling or emotion about the content of the image, or
   - The text only makes a reference to something shown in the image, or
   - The text is unrelated to the image

Examples for this task can be seen in Figure 1 by comparing Figures 1(a,c) (*Text is represented*) with Figures 1(b,d) (*Text is not represented*).

The second task – referred to as the **image task** in the rest of the paper – focuses on the role of the image to the semantics of the tweet and aims to identify if the image's content contributes with additional information to the meaning of the tweet beyond the text, as judged by an independent third party. This task is defined and annotated using the following guidelines:

1. Image has additional content that represents the meaning of the text and the image (**Image adds**):
   - Image contains other text that adds additional meaning to the text, or
   - Image depicts something that adds information to the text or
   - Image contains other entities that are referenced by the text.
2. Image does not add additional content that represents the meaning of text+image (**Image does not add**).

Examples for the image task can be seen in Figure 1 by comparing Figures 1(a,b) (*Image adds*) with Figures 1(c,d) (*Image does not add*).

Combining the labels of the two binary tasks described above gives rise to four types of text-image relationships (**Image+Text Task**). All of the four relationship types are exemplified in Figure 1.

4 Data Set

To study the relationship between the text and image in the same social media post, we define a new annotation schema and collect a new annotated corpus. To the best of our knowledge, no such corpus exists in prior research.

4.1 Data Sampling

We use Twitter as the source of our data, as this source contains a high level of expression of thoughts, opinions and emotions (Java et al., 2007; Kouloumpis et al., 2011). It represents a platform for observing written interactions and conversations between users (Ritter et al., 2010).

The tweets were deliberately randomly sampled tweets from a list of users for which several of their socio-demographic traits are known, introduced in past research (Preoţiuc-Pietro et al., 2017). This will enable us to explore if the frequency of posting tweets with a certain text-image relationship is different across socio-demographic groups.

We downloaded as many tweets as we could from these users using the Twitter API (up to 3,200 tweets/user per API limits). We decided to annotate only tweets from within the same time range (2016) in order to reduce the influence of potential platform usage changes with time. We filter out tweets that are not written in English using the langid.py tool (Lui and Baldwin, 2012).

In total, 2,263 users (out of the initial 4,132) have posted tweets with at least one image in the year 2016 and were included in our analysis. Our final data set contains 4,471 tweets.

4.2 Demographic Variables

The Twitter users from the data set we sampled have self-reported the following demographic variables through a survey: gender, age, education level and annual income. All users solicited for data collection were from the United States in order to limit cultural variation.

- Gender was considered binary[2] and coded with Female – 1 and Male – 0. All other variables are treated as ordinal variables.
- Age is represented as a integer value in the 13–90 year old interval.
- Education level is coded as an ordinal variable with 6 values representing the highest degree obtained, with the lowest being 'No high school degree' (coded as 1) and the highest being 'Advanced Degree (e.g., PhD)' (coded as 6).
- Income level is coded as on ordinal variable with 8 values representing the annual income of the person, ranging from '$<$ \$20,000' to '$>$ \$200,000').

For a full description of the user recruitment and quality control processes, we refer the interested reader to (Preoţiuc-Pietro et al., 2017).

4.3 Annotation

We have collected annotations for text-image pairs from 4,471 tweets using the Figure Eight platform (formerly CrowdFlower). We annotate all tweets containing both text and image using two independent annotation tasks in order to simplify the task and not to prime annotators use the outcome of one task as a indicator for the outcome of the other.

For quality control, 10% of annotations were test questions annotated by the authors. Annotators had to maintain a minimum accuracy on test questions of 85% for the image task and 75% for the text task for their annotations to be valid.

Inter-annotator agreement is measured using Krippendorf's Alpha. The overall Krippendorfs Alpha is 0.71 for the image task, which is in the upper part of the *substantial* agreement band (Artstein and Poesio, 2008). We collect 3 judgments and use majority vote to obtain the final label to further remove noise. For the text task, we collected and aggregated 5 judgments as the Krippendorf's Alpha is 0.46, which is considered moderate agreement (Artstein and Poesio, 2008). The latter task was more difficult due to requiring specific world knowledge (e.g. a singer mentioned in a text also present in an image) or contained information encoded in hashtags or usernames which the annotators sometimes overlooked. The aggregated judgments for each task were combined to obtain the four class labels. The label distributions of the aggregated annotations are: a) Text is represented & Image adds: **18.5%**; b) Text is represented & Image does not add: **21.9%**; c) Text is not represented & Image adds: **25.6%**; d) Text is not represented & Image does not add: **33.8%**.

[2]We asked users to report gender as either 'Female', 'Male' or an open-ended field, and removed the few users which did not select 'Male' or 'Female'

5 Methods

Our goal is to develop methods that are capable of automatically classifying the text-image relationship in tweets. We experiment with several methods which use information of four different types: demographics of the user posting the tweet, metadata from the tweet, the text of the tweet or the image of the tweet; plus a combination of them. The methods we use are described in this section.

5.1 User Demographics

User demographic features are the survey-based demographic information we have available for all users that posted the annotated tweets. The use of these traits is based on the intuition that different demographic groups have different posting preferences (Pennacchiotti and Popescu, 2011; Kosinski et al., 2013). We use this approach for comparison reasons only, as in practical use cases we would normally not have access to the author's demographic traits.

We code the gender, age, education level and income level of the user as features and use them in a logistic regression classifier to classify the text-image relationship.

5.2 Tweet Metadata

We experiment with using the tweet metadata as features. These code if a tweet is a reply, tweet, like or neither. We also add as features the tweet like count, the number of followers, friends and posts of the post's author and include them all in a logistic regression classifier.

These features are all available at tweet publishing time and we build a model using them to establish a more solid baseline for content based approaches.

5.3 Text-based Methods

We use the textual content of the tweet alone to build models for predicting the text-image relationship. We expect that certain textual cues will be specific to relationships even without considering the image content. For example, tweets ending in an ellipsis or short comments will likely be predictive of the text not being represented in the image.

Surface Features. We first use a range of surface features which capture more of the shallow stylistic content of the tweet. We extract number of tokens, uppercase tokens, exclamations, questions, ellipsis, hashtags, @ mentions, quotes and URLs from the tweet and use them as features in a logistic regression classifier.

Bag of Words. The most common approach for building a text-based model is using bag-of-words features. Here, we extract unigram and bigram features and use them in a logistic regression classifier with elastic net regularization (Zou and Hastie, 2005).

LSTM. Finally, based on recent results in text classification, we also experiment with a neural network approach which uses a Long-Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) network. The LSTM network processes the tweet sequentially, where each word is represented by its embedding ($E = 200$) followed by a dense hidden layer ($D = 64$) and by a a ReLU activation function and dropout (0.4) The model is trained by minimizing cross entropy using the Adam optimizer (Kingma and Ba, 2014). The network uses in-domain Twitter GloVe embeddings pre-trained on 2 billion tweets (Pennington et al., 2014).

5.4 Image-based Methods

We use the content of the tweet image alone to build models for predicting the text-image relationship. Similar to text, we expect that certain image content will be predictive of text-image relationships even without considering the text content. For example, images of people may be more likely to have in the text the names of those persons.

To analyze image content, we make use of large pre-trained neural networks for the task of object recognition on the ImageNet data set. ImageNet (Deng et al., 2009) is a visual database developed for object recognition research and consists of 1000 object types. In particular, we use the popular pre-trained InceptionNet model (Szegedy et al., 2015), which achieved the best performance at the ImageNet Large Scale Visual Recognition Challenge 2014 to build the following two image-based models.

ImageNet Classes. First, we represent each image in a tweet with the probability distribution over the 1,000 ImageNet classes obtained from InceptionNet. Then, we pass those features to a logistic regression classifier which is trained on our task. In this setup, the network parameters remain fixed, while only the ImageNet class weights are learned in the logistic regression classifier.

Tuned InceptionNet. Additionally, we tailored

the InceptionNet network to directly predict our tasks by using the multinomial logistic loss with softmax as the final layer for our task to replace the 1,000 ImageNet classes. Then, we loaded the pre-trained network from (Szegedy et al., 2015) and fine-tuned the final fully-connected layer with the modified loss layers. We perform this in order to directly predict our task, while also overcoming the necessity of re-extracting the entire model weights from our restricted set of images.

The two approaches to classification using image content based on pre-trained model on ImageNet have been used successfully in past research (Cinar et al., 2015).

5.5 Joint Text-Image Methods

Finally, we combine the textual and image information in a single model to classify the text-image relationship type, as we expect both types of content and their interaction to be useful to the task.

Ensemble. A simple method for combining the information from both modalities is to build an ensemble classifier. This is done with a logistic regression model with two features: the Bag of Words text model's predicted class probability and the Tuned InceptionNet model's predicted class probability. The parameters of the model are tuned by cross validation on the training data and similar splits as the individual models.

LSTM + InceptionNet. We also build a joint approach by concatenating the features from the final layers of our LSTM and InceptionNet models and passing them through a fully-connected (FC) feed forward neural network with one hidden layer (64 nodes). The final output is our text-image relationship type. We use the Adam optimizer to fine tune this network. The LSTM model has the same parameters as in the text-only approach, while the InceptionNet model is initialized with the pre-trained model on the ImageNet data set.

6 Predicting Text-Image Relationship

We split our data into a 80% train (3,576 tweets) and 20% test (895 tweets) stratified sample for all of our experiments. Parameters were tuned using 10-fold cross-validation with the training set, and results are reported on the test set. Table 1 presents the weighted F1-scores for the text task, the image task and the image+text task with all the methods described in Section 5. The weighted F1 score is the weighted average of the class-level F1 scores, where the weight is the number of items in each class.

Method	Image Task	Text Task	Image+Text Task
Majority Baseline	0.37	0.44	0.16
User Demographics	0.39	0.45	0.17
Tweet Metadata	0.38	0.48	0.21
Text-based Methods			
Surface Features	0.39	0.53	0.21
Bag of Words	0.56	0.56	0.33
LSTM	0.60	0.57	0.33
Image-based Methods			
ImageNet Classes	0.67	0.52	0.33
Tuned InceptionNet	0.76	0.53	0.39
Joint Text-Image Methods			
Ensemble	0.76	0.53	0.39
LSTM + InceptionNet	**0.81**	**0.58**	**0.44**

Table 1: Experimental results in predicting text-image relationship with different methods and grouped by modalities used in prediction. Results are presented in weighted F1 score.

The majority baseline always predicts the most frequent class in each task, namely: *Image does not add* for the image task, *Text is not represented* for the text task and *Image does not add & Text is not represented* for the Image + Text task.

The models using user demographics and tweet metadata show minor improvements over the majority class baseline for both tasks. When the two tasks are combined, both feature types offer only a slight increase over the baseline. This shows that user factors mildly impact the frequency with which relationship types are used, which will be explored further in the analysis section.

The models that use tweet text as features show consistent improvements over the baseline for all three tasks. The two models that use the tweet's topical content (Bag of Words and LSTM) obtain higher predictive performance over the surface features. Both content based models obtain relatively similar performance, with the LSTM performing better on the image task.

The models which use information extracted from the image alone also consistently outperform the baseline on all three tasks. Re-tuning the neural network performs substantially better than building a model directly from the ImageNet classes on the image task and narrowly outperforms the other method on the text task. This is somewhat expected, as the retuning is performed on this domain specific task.

When comparing text and image based models across tasks, we observe that using image features obtains substantially better performance on the image task, while the text models obtain bet-

ter performance on the text task. This is somewhat natural, as the focus of each annotation task is on one modality and methods relying on content from that modality are more predictive alone as to what ultimately represents the text-image relationship type.

Our naive ensemble approach does not yield substantially better results than the best performing methods using a single modality. However, by jointly modelling both modalities, we are able to obtain improvements – especially on the image task. This shows that both types of information and their interaction are important to this task. Methods that exploit more heavily the interaction and semantic similarity between the text and the image are left for future work.

We also observe that the predictive methods we described are better at classifying the image task. The analysis section below will allow us to uncover more about what type of content characterizes each relationship type.

7 Analysis

In this section, we aim to gain a better understanding of the type of content specific of the four text-image relationship types and about user type preferences in their usage.

7.1 User Analysis

Socio-demographic traits of the authors of posts are known to be correlated with several social media behaviors including text (Rao et al., 2010; Pennacchiotti and Popescu, 2011; Schwartz et al., 2013; Volkova et al., 2014; Lampos et al., 2014; Preoţiuc-Pietro et al., 2015a,b, 2016; Preoţiuc-Pietro et al., 2017; Preoţiuc-Pietro and Ungar, 2018) and images (Alowibdi et al., 2013; You et al., 2014; Farseev et al., 2015; Skowron et al., 2016; Liu et al., 2016; Guntuku et al., 2017; Samani et al., 2018; Guntuku et al., 2019). We hypothesize that socio-demographic traits also play a role in the types of text-image relationships employed on Twitter.

To measure this, we use partial Pearson correlation where the dependent variables are one of four socio-demographic traits described in Section 4.2. The independent variables indicate the average times with which the user employed a certain relationship type. We code this using six different variables: two representing the two broader tasks – the percentage of tweets where image adds information and the percentage of tweets where the text is represented in the image – and four encoding each combination between the two tasks.

In addition, for all analyses we consider gender and age as basic human traits and control for data skew by introducing both variables as controls in partial correlation, as done in prior work (Schwartz et al., 2013; Preoţiuc-Pietro et al., 2017; Holgate et al., 2018). When studying age and gender, we only use the other trait as the control. Because we are running several statistical tests at once (24) without predefined hypotheses, we use Bonferroni correction to counteract the problem of multiple comparisons. The results are presented in Table 2.

We observe that age is the only user demographic trait that is significantly correlated to text-image relationship preference after controlling for multiple comparisons and other demographic traits. The text-image relationship where the text is represented in the image, at least partially, is positively correlated with age (r = 0.117).

Further analyzing the four individual text-image relationship types reveals that older users especially prefer tweets where there is a semantic overlap between the concepts present in the text and the image, but the image contributes with additional information to the meaning of the tweet. This is arguably the most conventional usage of images, where they illustrate the text and provide more details than the text could.

Younger users prefer most tweets where the image adds information to the meaning of the tweet, but this has no semantic overlap with the text. These are usually tweets where the text represents merely a comment or a feeling expressed with the image providing the context. This represents a more image-centric approach to the meaning of the tweet that is specific to younger users. These correlations are controlled for gender.

Education was also correlated with images where the text was represented in the image ($r = 0.076$, $p < .01$, Bonferroni corrected), but this correlation did not meet the significance criteria when controlled for age to which education is moderately correlated ($r = 0.302$). This demonstrates the importance of controlling for such factors in this type of analysis. No effects were found with respect to gender or income.

Trait	Gender	Age	Education	Income
Image adds	-0.002	0.019	0.014	-0.020
Text represented	0.034	**0.117**	0.046	-0.016
Image does not add & Text not represented	-0.031	-0.061	-0.049	0.025
Image does not adds & Text represented	0.038	0.045	0.038	-0.004
Image adds & Text not represented	-0.004	**-0.070**	0.000	-0.009
Image adds Text represented	0.001	**0.095**	0.016	-0.015

Table 2: Pearson correlation between user demographic traits and usage of the different text-image relationship types. All correlations in bold are significant at $p < .01$, two-tailed t-test, **Bonferroni corrected** for multiple comparisons. Results for gender are controlled for age and vice versa. Results for education and income are controlled for age and gender.

7.2 Tweet Metadata Analysis

We adapt a similar approach to uncover potential relationships between the text-image relationship expressed in the tweet and tweet metadata features described in Section 5.2. However, after controlling for multiple comparisons, we are left with no significant correlations at $p < 0.01$ level. Hence, we refrain from presenting and discussing any results using this feature group as significant.

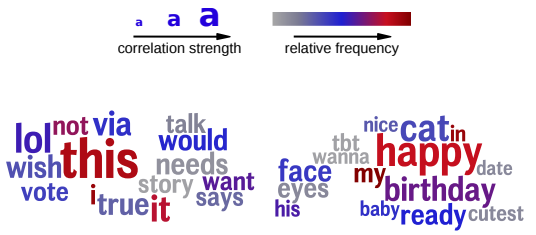
7.3 Text Analysis

Finally, we aim to identify the text and image features that characterize the four types of text-image relationship.

We use univariate Pearson correlation where the independent variable is each feature's normalized value in a tweet and the dependent variables are two binary indicators for the text and image tasks respectively. When performed using text features, this technique was coined Differential Language Analysis (Schwartz et al., 2013, 2017). The results when using unigrams as features are presented in Figure 3, 4 and 5.

Results for the image task (Figure 3) show that the image adds to the meaning of the tweet if words such as *this*, *it*, *why*, *needs* or *want* are used. These words can appear in texts with the role of referencing or pointing to an entity which is only present in the image.

Conversely, the image does not add to the meaning of the tweet when words indicative of objects that are also described in the image are present (*cat*, *baby*, *eyes* or *face*), thus resulting in the image not adding to the meaning of the tweet. A special case are tweets with birthday wishes, where a person is mentioned in text and also displayed in an image. Finally, the *tbt* keyword and hashtag is a popular social media trend where users post nostalgic pictures of their past accompanied by their textual description.



(a) Image adds (b) Image does not add

Figure 3: Words specific of each of the two classes from the image task when compared to the other.

(a) Text is represented (b) Text is not represented

Figure 4: Words specific of each of the two classes from the text task when compared to the other.

(a) Image does not add & Text not represented (b) Image does not add & Text represented

(c) Image adds & Text not represented (d) Image adds & Text represented

Figure 5: Words that are specific of each of the four classes compared to all other three classes. Font size is proportional to the Pearson correlation between each relationship type and word frequency. Color is proportional to the word frequency (see legend above the figures for reference).

The comparison between the two outcomes of the text task is presented in Figure 4. When the text and image semantically overlap, we observe words indicative of actions (*i've*), possessions (*your*) or qualitative statements (*congrats*, *loved*, *excited*, *tried*), usually about objects or persons also present in the image. We also observe a few nouns (*cats*, *hillary*) indicating frequent content that is also depicted in images (NB: the tweets were collected in 2016 when the U.S. presiden-

tial elections took place). Analyzing this outcome jointly with the text task, we uncover a prominent theme consisting of words describing first person actions (*congrats*, *thank*, *i've*, *saw*, *tell*) present when the image provides facets not covered by text (Figure 5d). Several keywords from text (*cat*, *game*, *winter*) show types of content which are present in both image and text, but the image is merely an illustrating these concepts without adding additional information (Figure 5a).

In contrast, the text is not represented in the image when it contains words specific of comments (*when*, *lmao*), questions (*do*, *was*), references (*this*) or ellipsis (*'...'*), all often referencing the content of the image as identified through data inspection. References to self, objects and personal states (*i*, *me*) and feelings (*miss*) are also expressed in text about items or things not appearing the image from the same tweet. Further exploring this result though the image task outcome, we see that the latter category of feelings about persons of objects (Figure 5a) – *miss*, *happy*, *lit*, *like*) are specific of when the image does not add additional information. Through manual inspection of these images, they often display a meme (as in Figure 1d) or unrelated expressions to the text's content. The image adds information when the text is not represented (Figure 5c) if the latter includes personal feelings, (*me*, *i*, *i'm*, *want*), comments (*lol*, *lmao*) and references (*this*, *it*), usually related to the image content as identified through an analysis of the data.

8 Conclusions

We defined and analyzed quantitatively and qualitatively the semantic relationships between the text and the image of the same tweet using a novel annotated data set. The frequency of use is influenced by the age of the poster, with younger users employing images with a more prominent role in the tweet, rather than just being redundant to the text or as a means of illustrating it. We studied the correlation between the content in the text and relation with the image, highlighting a differentiation between relationship types, even if only using the text of the tweet alone. We developed models that use both text and image features to classify the text-image relationship, with especially high performance (F1 = 0.81) in identifying if the image is redundant, which is immediately useful for downstream applications that maximize screen estate for users.

Future work will look deeper into using the similarity between the content of the text and image (Leong and Mihalcea, 2011), as the text task results showed room for improvements. We envision that our data, task and classifiers will be useful as a preprocessing step in collecting data for training large scale models for image captioning (Feng and Lapata, 2010) or tagging (Mahajan et al., 2018) or for improving recommendations (Chen et al., 2016) by filtering out tweets where the text and image have no semantic overlap or can enable new tasks such as identifying tweets that contain creative descriptions for images.

Acknowledgements

We like to thank our colleague Austin Ray for discussing the idea that originated this paper. We thank Ravneet Arora, Luka Bradesko, Prabhanjan Kambadur, Amanda Stent, Umut Topkara and the other members of the Bloomberg AI group who provided invaluable feedback on the experiments and paper. We also thank Eduardo Blanco for supporting the collaboration and feedback.

References

Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. CITE: A Corpus of Image-Text Discourse Relations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL, pages 570–575.

Malihe Alikhani and Matthew Stone. 2018. Exploring Coherence in Visual Explanations. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval, the First International Workshop on Multimedia Pragmatics*, MIPR, pages 272–277.

Malihe Alikhani and Matthew Stone. 2019. 'Caption' as a Coherence Relation: Evidence and Implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, NAACL, pages 58–67.

Jalal S. Alowibdi, Ugo A. Buy, and Philip Yu. 2013. Language Independent Gender Classification on Twitter. ASONAM.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Buffer. 2016. What 1 Million Tweets Taught Us About How People Tweet Successfully.

Tao Chen, Xiangnan He, and Min-Yen Kan. 2016. Context-aware Image Tweet Modelling and Recommendation. MM, pages 1018–1027.

Tao Chen, Dongyuan Lu, Min-Yen Kan, and Peng Cui. 2013. Understanding and classifying image tweets. MM, pages 781–784.

Tao Chen, Hany M SalahEldeen, Xiangnan He, Min-Yen Kan, and Dongyuan Lu. 2015. Velda: Relating an image tweet's text and images. AAAI, pages 30–36.

Yagmur Gizem Cinar, Susana Zoghbi, and Marie-Francine Moens. 2015. Inferring User Interests on Social Media From Text and Images. In *SoMeRa Workshop*, ICDM.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A Large-Scale Hierarchical Image Database. CVPR, pages 248–255.

Aleksandr Farseev, Liqiang Nie, Mohammad Akbari, and Tat-Seng Chua. 2015. Harvesting Multiple Sources for User Profile Learning: A Big Data Study. ICMR, pages 235–242.

Yansong Feng and Mirella Lapata. 2010. How many Words is a Picture Worth? Automatic Caption Generation for News Images. ACL, pages 1239–1249.

Sharath Chandra Guntuku, Weisi Lin, Jordan Carpenter, Wee Keong Ng, Lyle H Ungar, and Daniel Preoţiuc-Pietro. 2017. Studying Personality through the Content of Posted and Liked Images on Twitter. Web Science, pages 223–227.

Sharath Chandra Guntuku, Daniel Preotiuc-Pietro, Johannes C Eichstaedt, and Lyle H Ungar. 2019. What Twitter Profile and Posted Images Reveal About Depression and Anxiety. ICWSM.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural computation*, 9(8):1735–1780.

Eric Holgate, Isabel Cachola, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Why Swear? Analyzing and Inferring the Intentions of Vulgar Expressions. EMNLP, pages 4405–4414.

Mainak Jas and Devi Parikh. 2015. Image specificity. CVPR, pages 2727–2736.

Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private Traits and Attributes are Predictable from Digital Records of Human Behavior. *PNAS*, 110(15).

Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. 2011. Twitter Sentiment Analysis: The Good the Bad and the OMG! ICWSM, pages 538–541.

Vasileios Lampos, Nikolaos Aletras, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2014. Predicting and Characterising User Impact on Twitter. EACL, pages 405–413.

Chee Wee Leong and Rada Mihalcea. 2011. Measuring the semantic relatedness between words and images. In *Proceedings of the Ninth International Conference on Computational Semantics*, ACL, pages 185–194.

Leqi Liu, Daniel Preoţiuc-Pietro, Zahra Riahi, Mohsen E. Moghaddam, and Lyle Ungar. 2016. Analyzing Personality through Social Media Profile Picture Choice. ICWSM, pages 211–220.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. CVPR, pages 375–383.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf Language Identification Tool. ACL, pages 25–30.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. 2018. Exploring the Limits of Weakly Supervised Pretraining. ECCV, pages 185–201.

Emily E Marsh and Marilyn Domas White. 2003. A Taxonomy of Relationships between Images and Text. *Journal of Documentation*, 59(6):647–672.

Radan Martinec and Andrew Salway. 2005. A System for Image–Text Relations in New (and Old) Media. *Visual Communication*, 4(3):337–371.

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating Image Descriptions from Computer Vision Detections. EACL, pages 747–756.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal Named Entity Disambiguation for Noisy Social Media Posts. ACL, pages 2000–2008.

Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. 2011. Im2text: Describing images using 1 million captioned photographs. NIPS, pages 1143–1151.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A Machine Learning Approach to Twitter User Classification. ICWSM, pages 281–288.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. EMNLP, pages 1532–1543.

Daniel Preoţiuc-Pietro, Jordan Carpenter, Salvatore Giorgi, and Lyle Ungar. 2016. Studying the Dark Triad of Personality using Twitter Behavior. CIKM, pages 761–770.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015a. An Analysis of the User Occupational Class through Twitter Content. ACL, pages 1754–1764.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015b. Studying User Income through Language, Behaviour and Affect in Social Media. *PLoS ONE*.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond Binary Labels: Political Ideology Prediction of Twitter Users. ACL, pages 729–740.

Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. Developing User-Level Race and Ethnicity Predictors from Twitter Text. COLING, pages 1534–1545.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying Latent User Attributes in Twitter. SMUC.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised Modeling of Twitter Conversations. NAACL, pages 172–180.

Zahra Riahi Samani, Sharath Chandra Guntuku, Mohsen Ebrahimi Moghaddam, Daniel Preoţiuc-Pietro, and Lyle H Ungar. 2018. Cross-platform and Cross-interaction Study of User Personality based on Images on Twitter and Flickr. *PloS ONE*, 13(7).

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

H. Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Johannes Eichstaedt, and Lyle Ungar. 2017. Dlatk: Differential language analysis toolkit. EMNLP, pages 55–60.

Marcin Skowron, Marko Tkalčič, Bruce Ferwerda, and Markus Schedl. 2016. Fusing Social Media Cues: Personality Prediction from Twitter and Instagram. WWW Companion.

Ionut Sorodoc, Jey Han Lau, Nikolaos Aletras, and Timothy Baldwin. 2017. Multimodal topic labelling. volume 2 of *EACL*, pages 701–706.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper with Convolutions. CVPR, pages 1–9.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. CVPR, pages 3156–3164.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring User Political Preferences from Streaming Communications. ACL, pages 186–196.

Zhiyu Wang, Peng Cui, Lexing Xie, Wenwu Zhu, Yong Rui, and Shiqiang Yang. 2014. Bilateral Correspondence Model for Words-and-Pictures Association in Multimedia-Rich Microblogs. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 10(4):34:1–34:21.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1358–1368.

Quanzeng You, Sumit Bhatia, Tong Sun, and Jiebo Luo. 2014. The Eyes of the Beholder: Gender Prediction using Images Posted in Online Social Networks. ICDM.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Hui Zou and Trevor Hastie. 2005. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*.