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

Alakananda Vempala
Daniel Preoțiuc-Pietro
Motivation

What’s the largest difference in Twitter content in 2010 and 2019?
Motivation

What’s the largest difference in Twitter content in 2010 and 2019?

Many more tweets contain images
Motivation

What’s the largest difference in Twitter content in 2010 and 2019?

Many more tweets contain images

Approximately 12% tweets are now accompanied by images > 50M/day

Very little is known about the text-image relationship

TechAtBloomberg.com

© 2019 Bloomberg Finance L.P. All rights reserved.
Motivation

Text and image in a tweet can be related in several ways:

1. Text is the caption of the image
Motivation

Text and image in a tweet can be related in several ways:

2. Text is a comment on the image
Motivation

Text and image in a tweet can be related in several ways:

3. Image only illustrates part of the tweet text

HAPPY BIRTHDAY to our great @VP Mike Pence!
Applications

Automatically hiding images that don’t add additional content to a Twitter post to maximize screen estate.
Applications

• Automatically identify tweets that contain images and their captions

• Data can be used for distant supervision for image classification

• Identify tweets that contain creative descriptions for images
Aim: Determine whether there is semantic overlap between the context of the text and the image.

Some or all of the content words in the text are represented in the image (Text is represented).

None of the content words in the text are represented in the image (Text is not represented).
Data – Task Definition – Image Task

Focuses on the role of the image to the semantics of the tweet

Aim: Identify if the image’s content contributes with additional information to the meaning of the tweet beyond the text

<table>
<thead>
<tr>
<th>Image has additional content that represents the meaning of text &amp; image (Image adds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image does not add additional content that represents the meaning of text &amp; image (Image does not add)</td>
</tr>
</tbody>
</table>

---

**Bloomberg Economics** @economics · Jul 20
Africa’s potential workforce will top Asia’s by the end of this century as the continents wrestle with different demographic and economic challenges. bloom.bg/3ZQ3hqS

**Shifting Global Labor Pool**
What’s about to happen in one generation’s time

Note: Change in millions of population age 15-64 between 2020-2050
Source: Bloomberg analysis of the 2019 UN World Population Prospects

---

**Bloomberg LP** @Bloomberg · Dec 22, 2017
Through Bloomberg Startup, our employees help to grow strong global leaders of the future by collaborating closely with our non-profit education partners. bloom.bg/3F1vpxx

---

Bloomberg
Data – Annotation

• We used Figure Eight (formerly CrowdFlower)

• Text task
  — Redundancy of 3
  — Krippendorf’s Alpha = 0.71 (Text Task)
  — Annotators maintained > 85% accuracy over test questions

• Image task
  — Redundancy of 5
  — Krippendorf’s Alpha = 0.46 (Text Task)
  — Annotators maintained > 75% accuracy over test questions

• Adjudication by majority vote
Data – Collection

- Collected annotations for 4,888 tweets
  - All tweets posted in the same year (2016)
  - Split across original posts, retweets and favored posts
  - Deliberately sampled from users with known demographic traits
  - Tweets are all in English

- Available online: https://github.com/danielpreotiuc/text-image-relationship
Data – Distribution

Both task labels are combined to assign one of four classes to each text-image pair:

- **Image does not add & Text not represented**: 33.8%
- **Image does not add & Some text represented**: 21.9%
- **Image adds & Text not represented**: 25.6%
- **Image adds & Text is represented**: 18.5%
Analysis – Text Task

• Univariate Point-Biserial Correlation between unigram features and text task outcome
• Age is correlated with text being represented in image
  — Especially when image also adds information
  — More traditional type of relationship
• Simple tweet metadata not correlated

Some or all of the content words in the text are represented in the image (Text is represented)

None of the content words in the text are represented in the image (Text is not represented)
Analysis – Image Task

• Univariate Point-Biserial Correlation between unigram features and text task outcome
• Simple tweet metadata not correlated
• No demographic user information correlated
• 4-way analyses in the paper

Image does not add additional content that represents the meaning of text & image (Image does not add to meaning)
Prediction – Methods

- Demographics
- Metadata

- Text-Based Methods
  - Surface
  - Unigrams
  - BiLSTM

- Image-Based Methods
  - ImageNet Classes
  - InceptionNet Tuned

- Joint Text+Image approaches
  - Linear ensemble of text and image predictions
  - LSTM + InceptionNet architecture
Prediction – Baseline Methods

- Results in weighted F1, Train (80%), Test (20%), Parameters set via 10-fold CV
- Three tasks: Image Task (binary), Text Task (binary), Image + Text (4-class)
- Demographics, tweet metadata features almost no predictive value

![Graph showing F1 scores for Image Task, Text Task, and Image + Text Task]
Prediction – Text-based Methods

- All tasks show improvements on the baseline
  - Higher predictive power on image task (does the image have additional content)
- LSTM models are marginally better than unigrams

<table>
<thead>
<tr>
<th></th>
<th>Majority Class</th>
<th>Text-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Task</td>
<td>0.37</td>
<td>0.60</td>
</tr>
<tr>
<td>Text Task</td>
<td>0.44</td>
<td>0.57</td>
</tr>
<tr>
<td>Image + Text Task</td>
<td>0.16</td>
<td>0.33</td>
</tr>
</tbody>
</table>
Prediction – Image-based Methods

- All tasks show improvements on the baseline
  - Image-based methods > Text-based on the Image Task
  - Image-based methods < Text-based on the Text Task
- Tuned InceptionNet is the best image-based method

<table>
<thead>
<tr>
<th>Task Description</th>
<th>Image-based</th>
<th>Text-based</th>
<th>Image + Text-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Task (Image adds to meaning)</td>
<td>0.37</td>
<td>0.60</td>
<td>0.76</td>
</tr>
<tr>
<td>Text Task (Text is represented)</td>
<td>0.44</td>
<td>0.57</td>
<td>0.53</td>
</tr>
<tr>
<td>Majority Class</td>
<td>0.16</td>
<td>0.33</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Bar Chart: Comparison of performance metrics (e.g., accuracy) across different tasks and methods.
Prediction – Joint Text + Image Methods

- Improves over the best text-based or image-based methods
- LSTM + InceptionNet performs better
- Image task is much more predictable using models
- Text task is significantly harder (similar to humans)
Takeaways

Text-Image relationship in tweets is complex:
• Text does not always describe the image
• The image does not always illustrate text

Text-image relationship is likely useful for downstream applications

New classification schema and data set for text-image relationships on Twitter
• https://github.com/danielpreotiuc/text-image-relationship

Relationship type is predictable from both text and image
• Best results on each subtask are obtained by methods using different modalities (text or image)

We are hiring:
• NYC – http://careers.bloomberg.com/job/detail/74022