# PleaseUpVote: Virality Prediction with Title and Thumbnail Image on Reddit

Xiaowen Lin, Zijian Wang, Chaenan Ye

(vealin, zijiangwang, yec0014@stanford.edu)

## Introduction

Reddit is the 5th most visited website in the United States. A campaign or advertisement post can potentially reach millions of target audience. However, there is little work on predicting the virality, or popularity, of Reddit posts. This is indeed a hard task because even humans could hardly distinguish between two posts with different popularity. Here, we propose to train a multimodal neural network based on titles and thumbnails. Our results show that our model could capture the nuance signal of virality, and joining image and text information yields the best result.

## Data

We extracted four subreddits from Reddit dump from 2015 to 2018. Titles are included in the dump while thumbnails were crawled and resized to 224 x 224. Labels are derived from the upvotes of the posts. We split the dataset as 80% training, 10% dev, and 5% test.

<table>
<thead>
<tr>
<th>Subreddit</th>
<th>Number of Entries</th>
<th>Number of Avail. Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tech</td>
<td>4,573,934</td>
<td>1,988,266</td>
</tr>
<tr>
<td>EarthPorn</td>
<td>280,745</td>
<td>129,251</td>
</tr>
<tr>
<td>politics</td>
<td>1,432,923</td>
<td>545,644</td>
</tr>
<tr>
<td>AskSFM</td>
<td>1,720,414</td>
<td>1,976,371</td>
</tr>
</tbody>
</table>

Table 1: Sizes of datasets with numbers of available images

## Hyperparameters

- **Data**: Oversample or not
- **Image**: DenseNet / ResNet, with pre-training
- **Text embedding**: Word2Vec / Glove / Char / Non
- **Optimizer**: SGD w/ momentum, Adam or AMSGrad

## Models

![Diagram of models](image.png)

We investigate different state-of-the-art models in this project as follows:

- Image: DenseNet and ResNet with different sizes.
- Text: two-stack bi-directional LSTMs, with character-level and word-level embeddings.
- Multimodal: joining the feature layers of the pretrained models with fully-connected layers on top.

## Results

For the following section, we focus on ASFM as it has a large amount of data, and its relevance temporal. We formulate the task as a two-class classification problem, where we take the posts less than mean - std upvotes as negative, and more than mean + std as positive.

<table>
<thead>
<tr>
<th>Method</th>
<th>Macro-F1</th>
<th>ResNet + P + SGD</th>
<th>DenseNet + NF + SGD</th>
<th>DenseNet + P + SGD</th>
<th>DenseNet + P + Adam</th>
<th>LSTM + NF + SGD</th>
<th>LSTM + Glove + SGD</th>
<th>LSTM + Char + SGD</th>
<th>Multimodal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.678</td>
<td>0.653</td>
<td>0.762</td>
<td>0.681</td>
<td>0.618</td>
<td>0.617</td>
<td>0.609</td>
<td>0.738</td>
</tr>
</tbody>
</table>

Table 2: Performance comparison between different models.

## Discussions

- We demonstrate that there are sufficient cues to distinguish low and high-virality posts on some subreddits.
- The multimodal model performs the best, and images give stronger signals than textual data.
- Small vision models could capture virality signals well, while complicated ones do better.
- Using pretrained word embeddings does not lead to the better performance. The reason might be that the vocabulary on Reddit is different from the general text (Google News for Word2Vec, Wikipedia for Glove).

## Future Work

- Collect datasets with high resolution images: Thumbnails are small, but people may click into the post and view the original image.
- Model temporal subreddits: Our investigations show that it is very hard to predict virality in time-sensitive and news-related subreddits, e.g., politics and The_ Donald.
- Formulate the problem as a regression task: coming in the report.

## Acknowledgements

We thank Pedro Garzon, Ahmadreza Memari, and all CS230 staff for their suggestions and support through this project.

## References