Multimodal Sentiment Analysis for Image Text pair from Tweets

Team *Multi Mode*

Arkadip Maitra

Soumen Halder

arkadipmaitra@gmail.com
soumenhalder2018@gmail.com

Department of Computer Science
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, Howrah
Pin - 711 202, West Bengal

Abstract

Opinion and sentiment analysis is a vital task to characterize subjective information in social media posts. Multimodal sentiment analysis is a new dimension[peacock prose] of the traditional text-based sentiment analysis, which goes beyond the analysis of texts, and includes other modalities such as audio and visual data. With the extensive amount of social media data available online in different forms such as videos and images, the conventional text-based sentiment analysis has evolved into more complex models of multimodal sentiment analysis. In this paper, we present a comprehensive experiment for sentiment analysis of text and image pair from tweets. The main goal of this project is to show the ability of multi modal models to capture the sentiment from abundant social media data that are increasing every second.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Social media has become a phenomenon in terms of its usage by the general public, traditional media, enterprises, and also as a forum for discussing research in academia. With the evolution of the Internet, social media sites, in particular, have become multimodal in nature with content including text, audio, images, and videos to engage different senses of a user. Similarly, sentiment analysis techniques have also progressed from extensively explored textbased to multimodal sentiment analysis of image-text pairs or videos. With two or more modalities, the problem becomes more challenging since every modality might differently influence the overall sentiment, and modalities can have a complex interplay. For image-text pairs, this is even harder as images are perceived as a whole, whereas text is read sequentially.

Psychological studies have found that human visual attention generally prioritizes emotional content over non-emotional content. Some earlier other proposed models that prioritizes emotional objects over other objects to predict sentiment, just like human perception. It indicates that to learn a multimodal model for sentiment prediction, visual features should contribute and consider different objects, facial expressions, and other salient regions in the image. Besides, to learn a multimodal model for sentiment detection, extracted features from two modalities need to be combined in a way that reflects the overall sentiment of the image-text pair.

We use the publicly available benchmark MVSA [1] (Multi-View Social Data) that consists of different datasets of tweets and corresponding images. We provide a analysis of both image and text features to get the sentiment.

We created and used our test dataset in both English and Bengali language for evaluating the trained model.

This project is done following the paper 'A Fair and Comprehensive Comparison of Multimodal Tweet Sentiment Analysis Methods'[2] and we tried to get results close to this paper.

Chapter 2

Literature Review

Sentiment detection has been extensively explored for textual social media data, with earlier approaches that were lexicon-based evolving to statistical and machine learning-based classification in the last decade.
SentiStrength [3] is a well-known lexicon-based approach for short text built using widely occurring words and phrases on social media.
Later, Saif et al. [4] developed SentiCircles specifically for Twitter sentiment analysis by taking into account the co-occurrence of words in different contexts in tweets. With the prevalence of deep learning, convolutional neural networks (CNNs) and sequential models like Long Short-term Memory (LSTM) networks have been successfully used for tweet sentiment classification. Shin et al. [5] developed a hybrid approach by integrating lexicons with a CNN using an attention mechanism. With the rise of image and video data on social media sites like Instagram, Flickr, and Twitter, visual sentiment analysis has attracted a lot of attention recently.
Recently, Jiang et al. [6] proposed another attention mechanism where they used both cross-modal attention fusion followed by modality-specific CNN-gated feature extraction to learn a better representation. They used ImageNet pre-trained ResNet for visual features, and experimented with GloVe and BERT (Bidirectional Encoder Representations from Transformers) embeddings for textual features to achieve state-of-the-art results for the MVSA dataset.

Chapter 3

Methodology

In this section, we discuss the feature extraction and the steps that are used to preprocess and train the proposed multimodel encoder and classification.

3.1 Data Processing

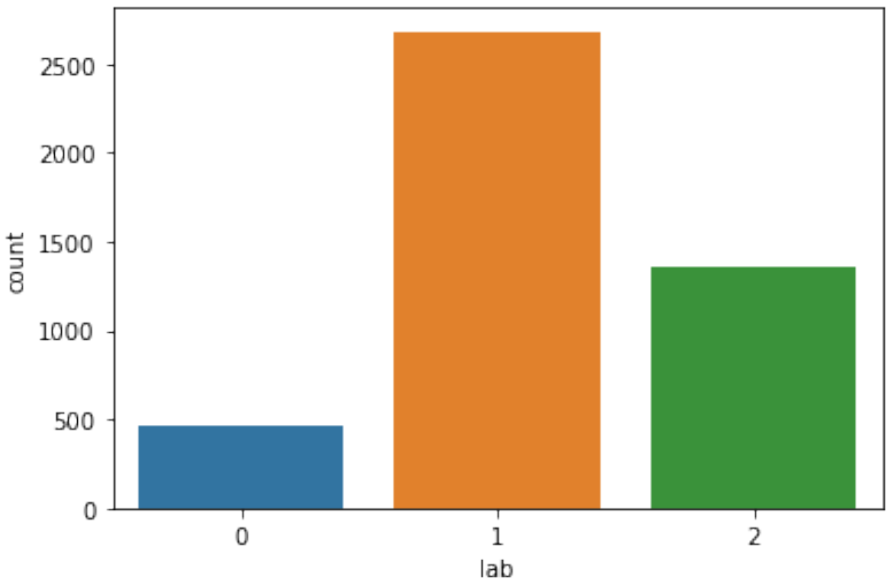


Figure 3.1: The distribution of data

The data consists of image file and text file for that image. A sentiment label of image and text is given separately. The sentiment class is of three types neutral, positive, negative. An aggregated sentiment is obtained by taking the non neutral

sentiment of the two for image text pair for conflict, the sentiment label of any one if no conflict and data is dropped if sentiment label has non neutral conflict.
This final data is then used for our experiment.
The data is not well distributed.

3.2 Visual Features

Visual features are extracted using a number of pretrained Resnet models. The different models are required to get features from different kinds of possible images.

3.2.1 Object Features

Different objects in a picture can incite a particular sentiment in a person. For instance, a cute dog or flowers might bring a positive sentiment, whereas a snake may incite a negative sentiment depending on the context. To encode objects and the overall image content, we extract features from a pre-trained ResNet model trained on the ImageNet dataset.
The final layer is a sequential layer of linear, batchnorm, relu to get a embedding of size 128.

3.2.2 Place and Scene Features

A scene or a place can also incite different sentiments in a person. For instance, a candy store might bring a positive sentiment, whereas a catacomb might incite a negative sentiment depending on the context. To encode the scene information of an image, we extract features from a pre-trained ResNet model trained on the Places365 dataset.
The final layer is a sequential layer of linear, batchnorm, relu to get a embedding of size 128.

3.2.3 Facial Expression Features

The presence of faces and facial expressions (smiling vs. sad face) in an image can also influence the sentiment in an observer. In the MVSA dataset, we found that around 50% of the images contain faces with an average of 2 to 3 faces per image. In order to encode information about facial expressions, we extract the final layer

features from a pre-trained Resnet model that is trained on face images based on the following seven classes: angry, disgust, fear, happy, sad, surprise and neutral.
The final layer is a sequential layer of linear, batchnorm, relu to get a embedding of size 128.

The embedding from all the image features are concated into a 3*128 dimensional embedding and passed through a sequential layer of linear, batchnorm, relu to get a embedding of size 128.

3.3 Textual Features

Since context and meaning of the words are equally important for the influence of the whole sentence towards the sentiment, we use Bert Cased to extract contextual word embeddings and employ different pooling strategies to get a single embedding for the tweet. We experimentally found that using Roberta gives similar result. Bert tokenizer was used for input encoding.
The processing of raw text is simple. We just encode the text in utf-8 and keep the case intatct. Hastag and url were seperated word by word to give the tokenizer ability to tokenize.
The final layer is a sequential layer of linear, batchnorm, relu and dropout to get a embedding of size 128.

3.4 Multimodal Features

The final embedding of text and image are of size 128. These are concated into 2*128 size tensor and passed through linear layer to get 3 length output that give the result of 3 classes of sentiment.

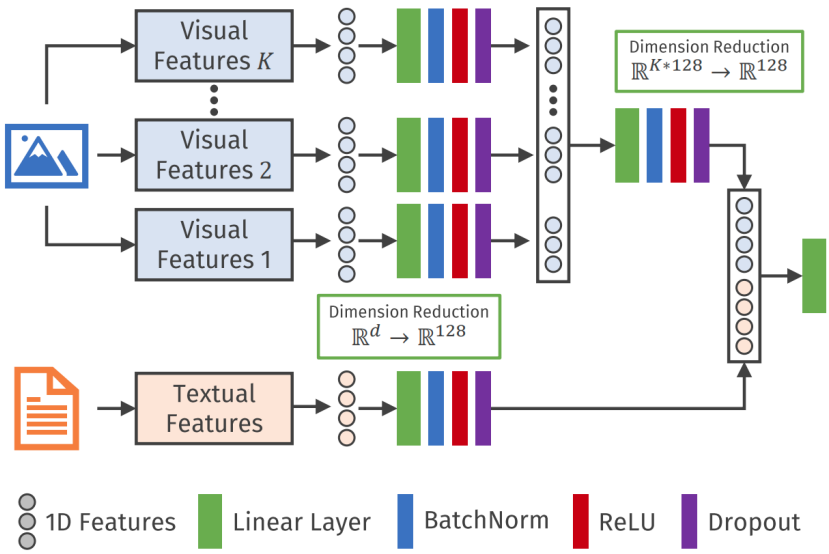


Figure 3.2: The Entire Model

The training is done only with the MVSA dataset which contains English language tweets only. The trained model is then tested with both of our English and Bengali dataset to evaluate performance .

Chapter 4

Experimental Results

The MVSA dataset is split into train and validation set where the validation set is 10% of the entire dataset. Our own created dataset is used as the test data.
The processing of the data is mentioned before.

4.1 Results of BERT-cased model

The entire dataset ie. the train data, validation data and test data is in English language.

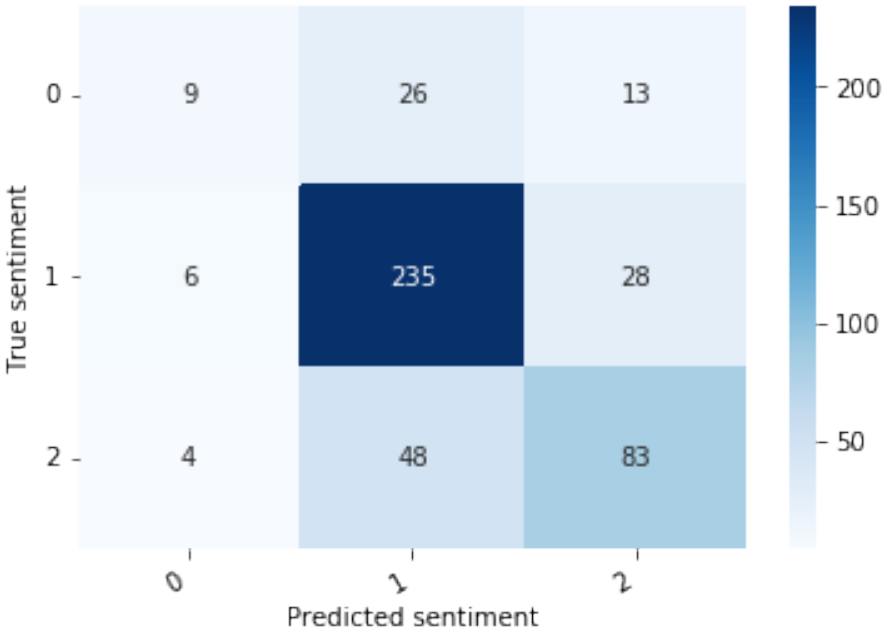
4.1.1 Results on Validation Set

	Precesion	Recall	F1-Score	Support
0	0.47	0.19	0.27	48
1	0.76	0.87	0.81	269
2	0.67	0.61	0.64	135
accuracy			0.72	452
macro-avg	0.63	0.56	0.57	452
weighted-avg	0.70	0.72	0.70	452

Table 4.1: Detailed Result of BERT model on Validation set

Highest Validation Accuracy = 72.34%

4.1.2 Confusion Matrix of Validation Set

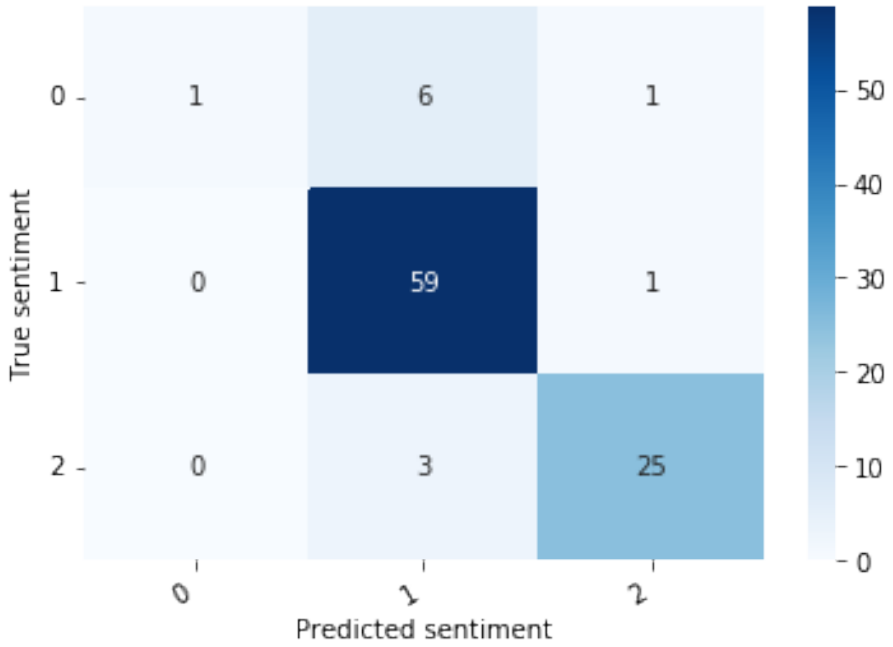


4.1.3 Results on English Test Set

	Precesion	Recall	F1-Score	Support
0	1	0.12	0.22	8
1	0.87	0.98	0.92	60
2	0.93	0.89	0.91	28
accuracy			0.89	96
macro-avg	0.93	0.67	0.68	96
weighted-avg	0.90	0.89	0.86	96

Table 4.2: Detailed Result of BERT model on English Test set

4.1.4 Confusion Matrix of Test Set



4.1.5 Comparison with State of the Art Model

Method	Max Accuracy	Max F1
MultiSentiNet[2]	69.25	63.61
FENet-BERT[2]	71.67	69.97
Se-MLNN[2]	82.04	81.14
Ours	87.62	86.0

Table 4.3: Comparison with SOTA

4.2 Results of Multilingual-BERT-cased model

The the train dataset, validation data are in English language but the test data is in Bengali language.

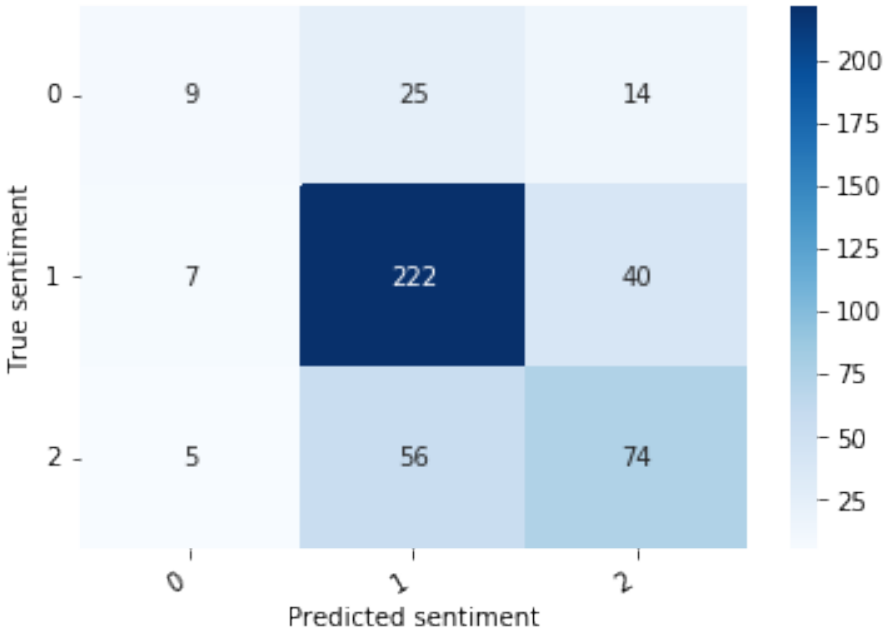
4.2.1 Results on Validation Set

	Precesion	Recall	F1-Score	Support
0	0.43	0.19	0.26	48
1	0.73	0.83	0.78	269
2	0.58	0.55	0.56	135
accuracy			0.67	452
macro-avg	0.58	0.52	0.53	452
weighted-avg	0.65	0.67	0.66	452

Table 4.4: Detailed Result of Multilingual-BERT model on Validation set

Highest Validation Accuracy = 67.47%

4.2.2 Confusion Matrix of Validation Set

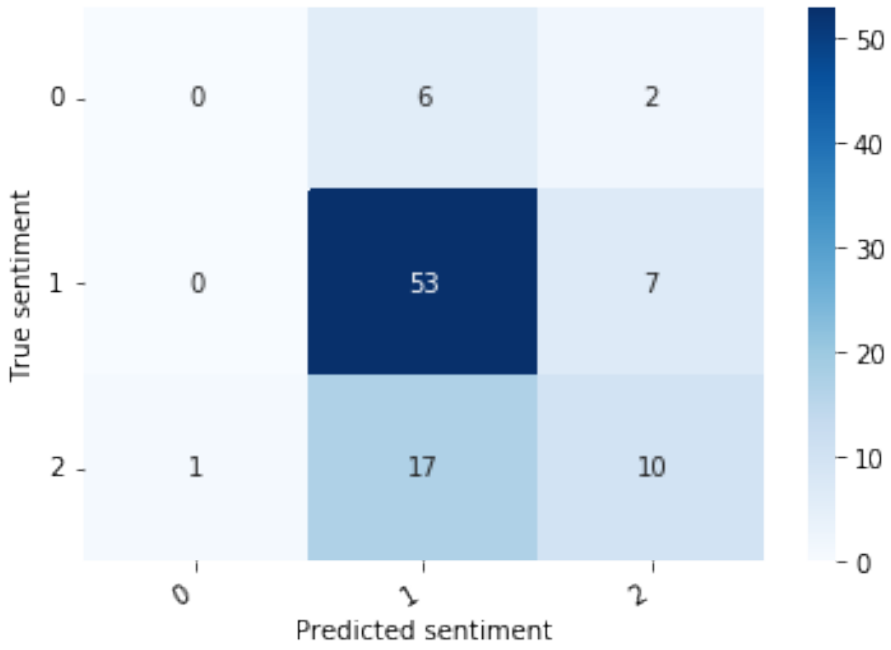


4.2.3 Results on English Test Set

	Precesion	Recall	F1-Score	Support
0	0	0	0	8
1	0.70	0.88	0.78	60
2	0.53	0.36	0.43	28
accuracy			0.66	96
macro-avg	0.41	0.41	0.40	96
weighted-avg	0.59	0.66	0.61	96

Table 4.5: Detailed Result of Multilingual-BERT model on English test set

4.2.4 Confusion Matrix of English Test Set



4.2.5 Comparison with State of the Art Model

Method	Max Accuracy	Max F1
MultiSentiNet[2]	69.25	63.61
FENet-BERT[2]	71.67	69.97
Se-MLNN[2]	82.04	81.14
Ours	64.94	66.0

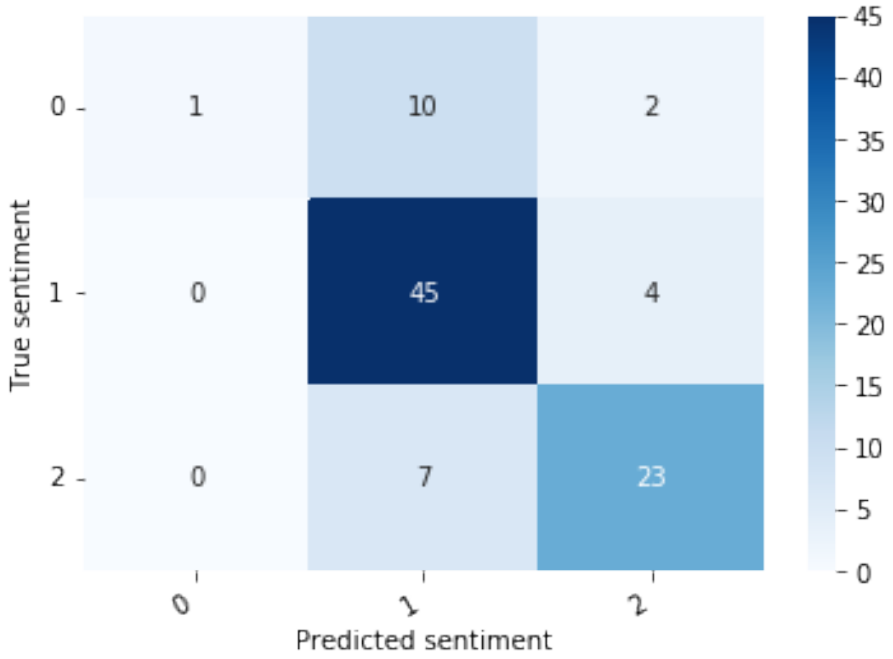
Table 4.6: Comparison with SOTA

4.2.6 Results on Bengali Test Set

	Precesion	Recall	F1-Score	Support
0	1	0.08	0.14	13
1	0.73	0.92	0.81	49
2	0.79	0.77	0.78	30
accuracy			0.75	92
macro-avg	0.84	0.59	0.58	92
weighted-avg	0.79	0.75	0.71	92

Table 4.7: Detailed Result of Multilingual-BERT model on Bengali Test set

4.2.7 Confusion Matrix of Bengali Test Set



Both the test dataset is created by us, as there is no standard test dataset. The data distribution is close to the original data.

Chapter 5

Summary

In summary, we have accomplished the followings:

- Experimentation with the MVSA dataset
- Training a multimodal model for Sentiment classification.
- Experimentation with various image and encoder models like Resnet50, Resnet101, Bert-Cased, Multilingual-BERT-cased etc.
- Creation of an English language test dataset for evaluation
- Making our project truly multilingual by showing the capability of our model on Bengali test dataset.
- Achieved accuracy metrics close to the paper for the English test set using the BERT-cased model

Chapter 6

References

[1] MULTIMEDIA COMMUNICATIONS RESEARCH LABORATORY. MVSA: Sentiment Analysis on Multi-view Social Data, https://mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social-data/

[2] Gullal S. Cheema, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth. 2021. A Fair and Comprehensive Comparison of Multimodal Tweet Sentiment Analysis Methods. In Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding (MMPT '21). Association for Computing Machinery, New York, NY, USA, 37–45. https://doi.org/10.1145/3463945.3469058

[3] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment in short strength detection informal text. J. Assoc. Inf. Sci. Technol. 61, 12 (2010), 2544–2558. https://doi.org/10.1002/asi.21416

[4] Hassan Saif, Yulan He, Miriam Fernández, and Harith Alani. 2016. Contextual semantics for sentiment analysis of Twitter. Inf. Process. Manag. 52, 1 (2016), 5–19. https://doi.org/10.1016/j.ipm.2015.01.005

[5] Bonggun Shin, Timothy Lee, and Jinho D. Choi. 2017. Lexicon Integrated CNN Models with Attention for Sentiment Analysis. In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2017, Copenhagen, Denmark, September 8, 2017, Alexandra Balahur, Saif M. Mohammad, and Erik van der Goot (Eds.). Association for Com-

putational Linguistics, 149–158. https://doi.org/10.18653/v1/w17-5220

[6] Tao Jiang, Jiahai Wang, Zhiyue Liu, and Yingbiao Ling. 2020. Fusion-Extraction Network for Multimodal Sentiment Analysis. In Advances in Knowledge Discovery and Data Mining - 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11-14, 2020, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12085), Hady W. Lauw, Raymond Chi-Wing Wong, Alexandros Ntoulas, Ee-Peng Lim, See-Kiong Ng, and Sinno Jialin Pan (Eds.). Springer, 785–797. https://doi.org/10.1007/978-3-030-47436-2_59