Text-Image Sentiment Analysis

Chen Qian[1], Edoardo Ragusa[2], Iti Chaturvedi[1],
Erik Cambria[1], and Rodolfo Zunino[2]

Nanyang Technological University, School of Computer Science and Engineering,
Nanyang 50 Ave, Singapore
Department of Electrical, Electronic, Telecommunications Engineering and Naval
Architecture, DITEN, University of Genoa, Genova, Italy

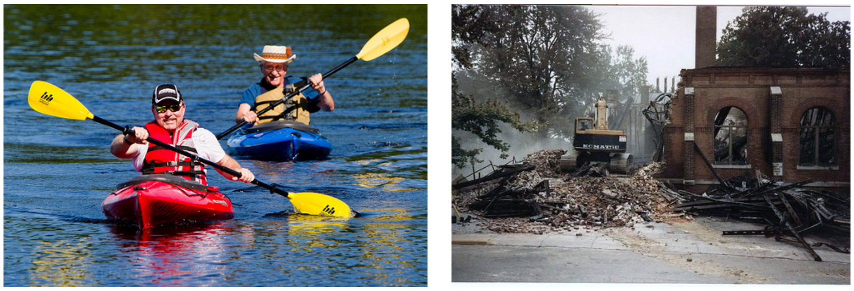
Abstract. Expressiveness varies from one person to another. Most images posted on Twitter lack good labels and the accompanying tweets have a lot of noise. Hence, in this paper we identify the contents and sentiments in images through the fusion of both image and text features. We leverage on the fact that AlexNet is a pre-trained model with great performance in image classification and the corresponding set of images are extracted from the web. In particular, we present a novel method to extract features from Twitter images and the corresponding labels or tweets using deep Convolutional Neural Networks (CNN) trained on Twitter data. We consider fine tuning AlexNet pre-trained CNN to initialize the model and AffectiveSpace of English concepts as text features. Lastly, to combine the image and text predictions we propose a novel sentiment score. Our model is evaluated on Twitter dataset of images and corresponding labels and tweets. We show that accuracy by merging scores from text and image models is higher than using any one system alone.

Keywords: Sentiment Analysis, Text-Image Joint, Weighted Score

1 Introduction

The proliferation of Web 2.0 technologies and the increasing use of computer-mediated communication resulted in exponential growth of information online. Despite the Internet's role as a facilitator of information in this Big Data Era, it has overloaded users with unrelated and noisy data. It is urgent to find an efficient and high-performance approach which extracts useful features from this massive amount of data. Natural Language Processing (NLP), a subdomain of Artificial Intelligence (AI), comes as a useful and practical method to handle the analysis of the language that humans use naturally in order to connect with computers and machines in both written and spoken contexts. For more than three decades, NLP has been handling problems between human and computer interaction.

NLP major tasks involve named entity recognition (NER), sentiment analysis, speech recognition, information retrieval, information extraction, relationship extraction, parsing, and machine translation, among others. Sentiment analysis,



(a) Positive

(b) Negative

Fig. 1. Examples for image polarities

one of its most interesting and challenging tasks, combines advanced techniques from NLP, machine learning, and information retrieval to extract opinions and subjective knowledge from online messages in social media. In fact, the rise and expansion of social media enabled millions of users to share their views, lives and interests in an impromptu manner and in real time, creating a huge amount of sentiment lexicons to be extracted and analyzed.

Initially, sentiment analysis focused on text documents such as product reviews and comments posted on social media platforms (e.g., Facebook, Twitter and Weibo). From the text, it was possible to extract the sentiment polarities of the sentences and classify them into positive, neutral and negative. It has been proposed that sentiment classifiers for text can be trained from positive and unlabeled examples using machine learning techniques [14] such as Naive Bayesian method and Support Vector Machines (SVM). Afterwards, sentiment analysis methods based on sentiment lexicons appeared and neural network enabled considerable progress in text sentiment classification.

Since the popularization of multimedia content in various social networks, text is no longer the only mode for sharing information. Image and videos are enabling people to express their thoughts and sentiments easily. As a consequence in this Big Data Era, sentiment analysis cannot be limited to text domain. In particular, images play a more important role in sentiment analysis, since they have the fullest quality of information. Furthermore, videos can also be represented as a sequence of images. For example, YouTube videos are a convenient way to share news events and product descriptions. Figure 1 (a) illustrates an image of two gentlemen paddling their canoes and laughing, annotated as positive polarity; and (b) illustrates an image of a building in ruins annotated as negative polarity.

CNN are the mainstream technique for image processing which can distinguish classes of images properly. Several authors have used CNN for object recognition and classification of images [10, 13, 23, 24]. It has been proved that Deep Convolutional networks were ideal for image sentiment analysis as it was able to detect over 10,000 different objects simultaneously. This powerful CNN architecture named AlexNet [13] is used in our method. Complementary to this, there are abundant Flickr datasets containing more than 3,000 Adjective Noun Pairs (ANPs), each image belongs to one ANP and the number of images in one ANP is ranging from dozens to thousands [10]. On the basis of ANPs, different levels of different emotions may be extracted [4, 23].

In this paper, we propose a method using textual and visual features to predict the sentiment polarity of Tweets containing both image and text. Following is the structure of this paper: Section 2 introduces studies related to sentiment analysis and the state-of-the-art methods; Section 3 is the preliminaries for our approach; Section 4 states our methods in a detailed way; Section 5 is the results and evaluation for our methods; finally, we summarize our work in Section 6.

2 Related Work

Traditional methods for sentiment analysis are mainly applied to text mining, which do not consider the presence of multimodal data, e.g., videos or images. As one of popular data format, images present more information but are more complex in compare to text.

As a novel and widely applied branch of NLP, sentiment analysis is to analyze the sentiment polarity of data, namely the attitudes, emotions and opinions of the data, which can be applied in marketing decision and product optimization for companies, as well as purchase decision for individual customers. A wheel of emotions created by R. Plutchik sorts emotions into 8 primary bipolar emotions and makes emotions in a colorful wheel showing the connection between emotions and colors [18]. Plutchik's Wheel of Emotions provides the basis for sentiment analysis, making it easier to handle nuanced semantic concepts and intuitively presenting sentiments in term of visual perception. Cambria et al. proposed the hourglass model inspiring by Plutchik's studies [8]. Roughly we divide sentiments into three polarities: positive, neutral and negative. In the 3D hourglass model, affective states from strongly positive to null to strongly negative are shaped to an hourglass containing four basic emotions: Attention, Aptitude, Pleasantness and Sensitivity. It uses basic emotions pairwise to result complex emotions.

ConceptNet [15] is a representation of the Open Mind Common Sense corpus representing noun phrases, verb phrases, adjective phrases or prepositional phrases by concept nodes. WordNet-Affect is a linguistic resource for the lexical representation of affective knowledge containing 1,903 terms referring to mental [21]. By applying the blending technique on ConceptNet and WordNet-Affect, Cambria et al. developed AffectiveSpace, a suitable knowledge base for emotive reasoning [7]. Blending Common Sense and Affective Knowledge to Perform Emotive Reasoning maps the most common relations in the affective knowledge base into ConceptNets set of relations. The latest version AffectiveSpace 2 improved by the Hourglass of Emotions contains 50,000 concepts with 100-dimension features for each concept, which can be embedded in potentially any cognitive system dealing with real-world semantics [6]. AffectiveSpace 2 generally outperforms standard AffectiveSpace, especially for categories where polarity is more difficult to detect in which affect is usually conveyed more implicitly.

In 2012, Siersdorfer et al. predicted sentiment of images using color histograms and Scale-Invariant Feature Transform (SIFT) techniques dataset with more than half a million Flickr images [20]. They used SentiWordNet as query terms to gather images with sentiment orientations. The bag-of-visual words representation and the color distribution of images are used to learn the image features. Through studying the connection between sentiment of images expressed in metadata and their visual content, Siersdorfer et al. achieved the precision values of up to 70% but with low recall values. Zhang et al. processed Sentiment Analysis on Microblogging by integrating text and image features [25]. In 2016, Katsurai et al. proposed a method mapping visual, textual and sentiment views into the latent embedding space and using correlations among these features [12]. The visual features is learnt from color histogram of images and this method achieved an accuracy of 74.77% on Flickr dataset and 73.60% on Instagram dataset.

DeepSentiBank [10] containing more than 3,000 ANPs significantly improved in both annotation accuracy and retrieval performance [17], compared to its predecessors which mainly use binary SVM classification models. Based on neural network, CNN introduced convonlutional filters to extract features and obtained outstanding achievements in image processing and deep learning. You et al. proposed a progressive CNN architecture [24] on CAFFE [11]. They trained half a million samples with ANPs from Flickr and fine-tuned the deep network using a progressive strategy therefore obtained a considerable accuracy with high recall. You et al. proposed to use CNN for the extraction of visual features and made fusions with textual features extracted from an unsupervised language model by learning distributed representations for documents and paragraphs [23]. Their model achieved a precision of 0.776 with recall of 0.740 by Early Fusion. Campos et al. provided a deep-dive analysis into *CaffeNet* and presented several experiments studying for the task of visual sentiment prediction [9].

3 Preliminaries

In this section, we will give the theoretical basis about CNN and AffectiveSpace 2 for our method.

3.1 Deep Convolutional Neural Network

CNN are a specific class of neural networks, based on four main building blocks: convolutions (kernels), non-linearities, pooling and dropout layers. A CNN is comprised of one or more convolutional layers (kernels) alternated by non linearities. Between them are inserted pooling and dropout layers. Finally, one or more fully connected layers as in a standard neural network gives the classification results.

Each convolutional layer works as a feature extractor. Using objects recognition as an example, lower level detects simple features like straight edges, simple colors, and curves, higher level extracts more complex features like noses, eyes.

Typical non linearities are *ReLu* and *Tanh*. Dropout layer are a regularization technique for reducing overfitting in neural networks by preventing complex co-adaptation on training data. Max pooling layers performs down-sampling by dividing the input into pooling regions, and computing the maximum of each region. The fully connected layer merges the extracted feature in order to perform the classification task.

This models present a huge number of parameters and are trained using standard back propagation techniques. Training from scratch requires labeled datasets with millions of patterns. For this reason in many applications transfer learning is applied. Transfer learning consist in remove the last fully-connected layer from a fully trained CNN, and replacing it with a new one. Then fine-tuning is applied to the weights of the new network by continuing the back -propagation. The main advantage of this technique is that it is possible to exploit feature detector for similar tasks, as an example adapt features for object recognition to polarity detection.

3.2 AffectiveSpace 2

To improve our model with textual features, we projected textual tweets data to AffectiveSpace 2. It is an effective way to cope with the evergrowing number of concepts and semantic features using AffectiveSpace. Cambria et al. replaced singular value decomposition (SVD), a low-rank approximation method, with random projection (RP) [3] to map the original high-dimensional data-set into a much lower-dimensional subspace by using a Gaussian N(0, 1) matrix, while preserving the pair-wise distances with high probability. This follows Johnson and Lindenstrausss (JL) Lemma [2]. The JL Lemma states that with high probability, for all pairs of points $x, y \in X$ simultaneously, there is:

$$\sqrt{\frac{m}{d}}\|x-y\|_2\,(1-\varepsilon) \leq \|\Phi_x - \Phi_y\|_2 \leq \sqrt{\frac{m}{d}}\,\|x-y\|_2\,(1+\varepsilon) \tag{1}$$

where X is a set of vectors in Euclidean space, d is the original dimension of this Euclidean space, m is the dimension of the space we wish to reduce the data points to, ε is a tolerance parameter measuring to what extent is the maximum allowed distortion rate of the metric space, and Φ is a random matrix.

Sarlos introduced that structured random projection for making matrix multiplication much faster [19]. Achlioptas and Dimitris proposed sparse random projection [1] to replace the Gaussian matrix with i.i.d. entries in:

$$\phi_{ji} = \sqrt{s}\begin{cases} 1 & with\, prab \cdot \dfrac{1}{2s} \\ 0 & with\, prob.1 - \dfrac{1}{s} \\ -1 & with\, prob. \dfrac{1}{2s} \end{cases} \tag{2}$$

where one can achieve a ×3 speedup by setting s = 3, since only one third of the data need to be processed.

When the number of features is much larger than the number of training samples ($d \gg n$), subsampled randomized Hadamard transform (SRHT) is preferred, as it behaves very much like Gaussian random matrices but accelerates the process from $O(nd)$ to $O(nlogd)$ time [16]. From [22, 16], for $d = 2^p$ where p is any positive integer, a SRHT can be defined as:

$$\phi = \sqrt{\frac{d}{m}} RHD \tag{3}$$

where m is the number we want to subsample from d features randomly; R is a random $m \times d$ matrix (the rows of R are m uniform samples from the standard basis of $\mathbb{R}^d$); $H \in \mathbb{R}^{d \times d}$ is a normalized Walsh-Hadamard matrix, which is defined recursively: $H_d = \begin{bmatrix} H_{k/2} & H_{k/2} \\ H_{k/2} & H_{k/2} \end{bmatrix}$ with $H_2 = \begin{bmatrix} +1 & +1 \\ +1 & 1 \end{bmatrix}$ and D is a $d \times d$ diagonal matrix and the diagonal elements are i.i.d. Rademacher random variables.

AffectiveSpace 2 is a vector space model preserving the semantic and affective relatedness of common-sense concepts while being highly scalable. In our method, it is an important part to extract sentiment features from textual Twitter data using AffectiveSpace 2.

4 Proposed Framework

This paper proposed approach consists in merging the information from images and their captions. The feature from the image and the caption are extracted independently. The text features are extracted by means of single world projections on affectspace 2. The visual feature are extracted using AlexNet and fine-tuned by Twitter images.

4.1 Visual features

The proposed feature extraction model of the CNN is inspired from AlexNet. Since AlexNet is a deep CNN architecture trained on 1.2 million images for the task of object detection with considerable precisions, it is efficient to detect sentiment of images by fine-tuning AlexNet with labeled images. In our model, we removed the fully connected layers and replaced it with a fully connected layer of size 4096 × 2. Then we fine tune the weights using about 18928 pattern from Twitter, derived from 301 negative images and 581 positive images.

4.2 Textual features

The textual features are extracted by 5-fold cross-validation method with AffectiveSpace. The original samples are randomly partitioned into 5 equal sized subsamples. Of the 5 subsamples, we train four of them and the other subsample is retained as the validation data. This process is repeated 5 times. Supporting Vector Machines (SVM) is used in the procedure of extracting textual features.

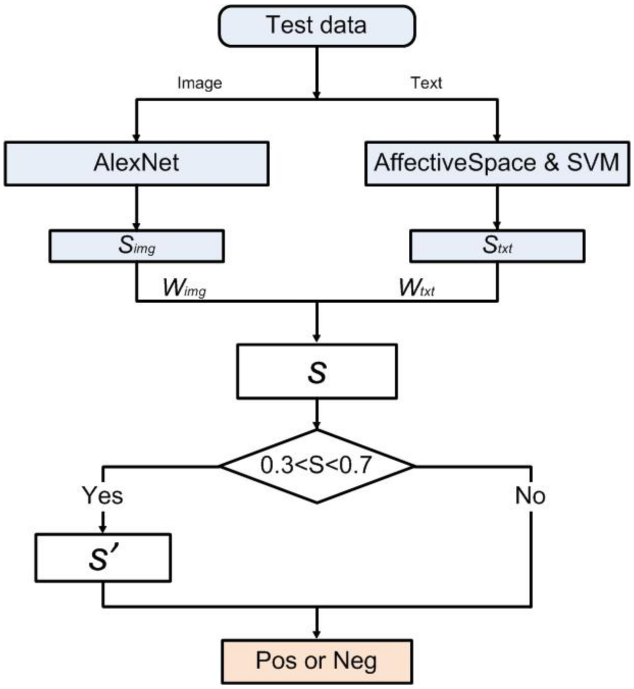


Fig. 2. Framework for TISC

4.3 TISC Model

Sentiment polarity prediction is based on features extracted from images and texts respectively. In order to balance visual features and textual features, our Text-Image Sentiment Classification (TISC) model employs the following computational approach to obtain sentiment scores of test data. Fig. 2 is the flowchart of computing sentiment scores. First of all, we define the preliminary weight of our image features extraction architecture as:

$$w_{img} = \frac{Acc_{img}}{Acc_{img} + Acc_{txt}} \tag{4}$$

where Acc_{img} and Acc_{txt} is the accuracy for validation of image and text data respectively. Similarly, the textual weight is $w_{txt} = 1 - w_{img}$.

The preliminary sentiment score s is calculated by the following equation.

$$s = s_{img}w_{img} + s_{txt}w_{txt} \tag{5}$$

where s_{img} is the sentiment score for test image predicted by CNN, while s_{txt} is the score for text data given from AffectiveSpace 2.

For test data with $s \in [0.3, 0.7]$, we assume that such data do not have strong sentiment polarities or there is conflict between textual and visual features.

Hence, using the weighted scores to classify the sentiment polarities of these data is not appropriate. We import a new measure s' with variables $\alpha, \beta \in [0,1]$ giving weights to visual and textual system respectively:

$$s' = 1 - (\alpha s_{img} + \beta s_{txt}) \tag{6}$$

With the sentiment scores measured as above, we define the sentiment polarity with 1 for *Positive* and 0 for *Negative* calculated by the following equation:

$$Polarity = \begin{cases} 1 & with\ s \geq 0.7\ \ or\ \ s' \geq 0.5 \\ 0 & with\ s \leq 0.3\ \ or\ \ s' < 0.5 \end{cases} \tag{7}$$

5 Experiments and evaluation

In this section we first introduce the feature extraction from images for sentiment classification. Next, we compare the accuracy of TISC model with baselines for image sentiment analysis.

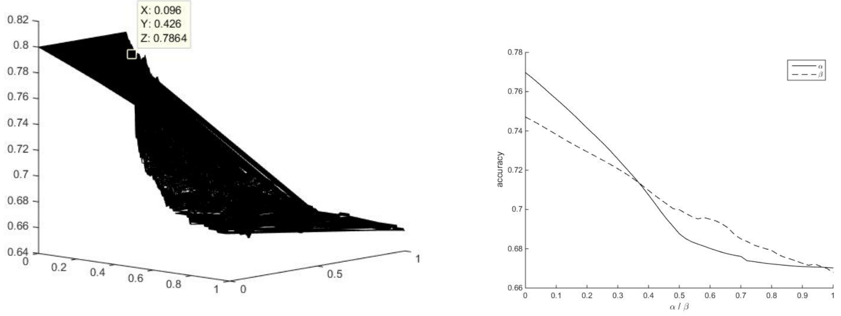
5.1 Datasets

To conduct experiments on textual and visual features, we fine-tune AlexNet with 1269 labeled Twitter images[1]. These images are annotated by 5 Amazon Mechanical Turk workers and we choose the images with the same sentiment label given by all the 5 workers (581 positive and 301 negative images). Since the data are unbalanced, in order to improve the fine-tuning on AlexNet, we double the negative images and increase the size of dataset to 18928 by adding rotations and reversals of each image. To judge our model, we test it on 596 images (463 positive and 133 negative images) with captions from Twitter [5].

5.2 Experiment Results

In the first step of our approach, we fine tune the AlexNet with Flickr dataset to obtain the visual kernel features. Next, textual features are extracted by 5-fold validation using SVM classifier. To find the optimal combination of textual features and visual features, we use trial and error method, we progressively change all the parameters $\alpha, \beta \in [0,1]$ in step of 0.002 (251,001 pairs of α and β). The highest accuracy in Fig. 3(a) reached 0.8051 and the accuracy is stable in the left part. Fig. 3(b) shows how the value of α or β effects the accuracy and for each curve, the accuracy of α is the average accuracy for β on each value of α, which is similar to β. Table 1 shows the results of sentiment prediction using different methods and TISC using diverse parameters.

[1] http://www.cs.rochester.edu/u/qyou/DeepSent/deepsentiment.html



(a) Accuracy of Text-Image

(b) Accuracy - α/β curve

Fig. 3. Accuracy trend with α and β

Table 1. Results of different methods

Method	α	β	Pred_neg	Pred_pos	Rec_neg	Rec_pos	Accuracy
Visual feature	-	-	0.3878	0.8352	0.4385	0.8043	0.7169
Textual feature	-	-	0.4122	0.8439	0.4692	0.8109	0.7356
TISC(1)	0.054	0.448	**0.6596**	0.8177	0.2385	0.9652	**0.8051**
TISC(2)	0.284	0.244	0.5574	0.8185	0.2615	0.9413	0.7915
TISC(3)	0.030	0.578	0.4787	0.8286	0.3462	0.8935	0.7729
TISC(4)	0.300	0.340	0.4058	0.8371	0.4308	0.8217	0.7356
TISC(5)	0.792	0.798	0.3610	0.8768	**0.6692**	0.6652	0.6661

Table 2. Comparison of other methods evaluated by AUC

Method	AUC
Low-level Features [5]	0.528
SentiBank [4]	0.514
TISC	**0.586**

5.3 Evaluation

Fig. 3 shows the accuracy of TISC corresponding to different α and β. For example, the accuracy achieves 0.7864 at point $(0.096, 0.426)$. From Fig. 3(a), it is shown that for α, β, if $\alpha + \beta \in (0, 0.5)$, then the results of system are stable with accuracy of 0.8051. However, from eqn (6) reveals the truth that a high accuracy is with a low recall for Negative data since when $\alpha + \beta < 0.5$, for all data there is $s' > 0.5$ and then be classified as positive data. Fig. 3(b) is the trend curves for accuracy-α/β from where we can see that in [0,1], the higher the values of α/β is, the lower the accuracy is.

From Table. 1, the last five rows are the results reflecting to different (α, β) pairs with $\alpha + \beta > 0.5$, we can find that the TISC has a significant increase in sentiment prediction compared with using single measures to predict sentiment polarity of test data and precision and recall are negative correlated. Table 2 shows that our model outperforms baselines by almost 6% in AUC. We compare with two baselines: Low-level Features are a set of features that can be useful for characterizing sentiment clues such as scenes, textures, and faces as well as other abstract concepts [5]; SentiBank is a concept representation with detectors trained on Flickr images.

6 Conclusion

This paper considers the application of Twitter images with captions for the prediction of sentiments applying fine-tune techniques. The Twitter images have corresponding labels or tweets, hence the merging of features from images and text is proposed. In this way, we can predict image sentiment as positive or negative with better performance. We see that the accuracy after fusing text and image features is higher than using a single modality. To extract the image features we consider AlexNet, which is a previously trained deep convolutional architecture. For text features we extract the significant concepts and project them on the AffectiveSpace of emotions. Lastly, we propose a novel sentiment scores to combine the prediction from image and text features.

References

1. Achlioptas, D.: Database-friendly random projections: Johnson-lindenstrauss with binary coins. Journal of computer and System Sciences **66**(4) (2003) 671–687
2. Balduzzi, D.: Randomized co-training: from cortical neurons to machine learning and back again. arXiv preprint arXiv:1310.6536 (2013)
3. Bingham, E., Mannila, H.: Random projection in dimensionality reduction: applications to image and text data. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2001) 245–250
4. Borth, D., Chen, T., Ji, R., Chang, S.F.: Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In: Proceedings of the 21st ACM international conference on Multimedia, ACM (2013) 459–460
5. Borth, D., Ji, R., Chen, T., Breuel, T., Chang, S.F.: Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: Proceedings of the 21st ACM international conference on Multimedia, ACM (2013) 223–232
6. Cambria, E., Fu, J., Bisio, F., Poria, S.: Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis. In: AAAI. (2015) 508–514
7. Cambria, E., Hussain, A., Havasi, C., Eckl, C.: Affectivespace: Blending common sense and affective knowledge to perform emotive reasoning. WOMSA at CAEPIA, Seville (2009) 32–41
8. Cambria, E., Livingstone, A., Hussain, A.: The hourglass of emotions. Cognitive behavioural systems (2012) 144–157
9. Campos, V., Salvador, A., Giro-i Nieto, X., Jou, B.: Diving deep into sentiment: Understanding fine-tuned cnns for visual sentiment prediction. In: Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia, ACM (2015) 57–62

10. Chen, T., Borth, D., Darrell, T., Chang, S.F.: Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. arXiv preprint arXiv:1410.8586 (2014)
11. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: tional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia, ACM (2014) 675–678
12. Katsurai, M., Satoh, S.: Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, IEEE (2016) 2837–2841
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
14. Liu, B., Dai, Y., Li, X., Lee, W.S., Yu, P.S.: Building text classifiers using positive and unlabeled examples. In: Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, IEEE (2003) 179–186
15. Liu, H., Singh, P.: Conceptneta practical commonsense reasoning tool-kit. BT technology journal **22**(4) (2004) 211–226
16. Lu, Y., Dhillon, P., Foster, D.P., Ungar, L.: Faster ridge regression via the subsampled randomized hadamard transform. In: Advances in neural information processing systems. (2013) 369–377
17. Narihira, T., Borth, D., Yu, S.X., Ni, K., Darrell, T.: Mapping images to sentiment adjective noun pairs with factorized neural nets. arXiv preprint arXiv:1511.06838 (2015)
18. Plutchik, R.: Plutchiks wheel of emotions (1980)
19. Sarlos, T.: Improved approximation algorithms for large matrices via random projections. In: Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on, IEEE (2006) 143–152
20. Siersdorfer, S., Minack, E., Deng, F., Hare, J.: Analyzing and predicting sentiment of images on the social web. In: Proceedings of the 18th ACM international conference on Multimedia, ACM (2010) 715–718
21. Strapparava, C., Valitutti, A.: Wordnet affect: an affective extension of wordnet. In: LREC. Volume 4. (2004) 1083–1086
22. Tropp, J.A.: Improved analysis of the subsampled randomized hadamard transform. Advances in Adaptive Data Analysis **3**(01n02) (2011) 115–126
23. You, Q., Luo, J., Jin, H., Yang, J.: Joint visual-textual sentiment analysis with deep neural networks. In: Proceedings of the 23rd ACM international conference on Multimedia, ACM (2015) 1071–1074
24. You, Q., Luo, J., Jin, H., Yang, J.: Robust image sentiment analysis using progressively trained and domain transferred deep networks. In: AAAI. (2015) 381–388
25. Zhang, Y., Shang, L., Jia, X.: Sentiment analysis on microblogging by integrating text and image features. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer (2015) 52–63