Human-Inspired Computer Vision & Reasoning
9-07

Group Leader: Ho Jun Yi (4i1)

Members: Joey Tang (4P3), Lai Jun Yu (4P3)

1 Introduction & Rationale

1a Idea Description

Image sentiment analysis is fast becoming a quintessential tool for various parties to classify and process the wealth of images available online, especially on social media. Seeing as images posted on various social media platforms such as Instagram, Twitter and Facebook are commonly accompanied by a text description, we propose that textual sentiment analysis of the various descriptors can significantly improve the reliability and accuracy of an image sentiment analysis model through the regression of multi-media vectors.

1b Significance and uses

With the increasing popularity of social media platforms globally, the massive amount of data they generate cannot be overlooked. Sentiment analysis of various posts has a wide range of applications, from publicity to social media monitoring. Therefore, by accurately and reliably predicting the sentiment of social media posts, and by extension - the general sentiment of the public, organizations will be able to better cater to the interests of their target audience (eg. product analysis).

2 Literature review

a. Cross-Media Learning for Image Sentiment Analysis in the Wild[1]

Due to a lack of extensive, labelled datasets for sentiment analysis, this approach uses textual analysis to exploit large-scale datasets of unlabelled images, intended for unsupervised image sentiment analysis. Text extracted from Twitter was classified according to sentiment polarity using the ItaliaNLP Sentiment Polarity Classifier[3], which is based on a tandem LSTM-SVM architecture. This architecture uses the LSTM networks to learn long-term order dependencies, and the SVM classification algorithms to identify sporadic textual features. Data with the most confident textual sentiment predictions was selected and sentiment labels were assigned to the corresponding images. Successful Deep CNNs were pre-trained on generic image datasets and fine-tuned with the labelled data from textual analysis. The best model (VGG-T4SA FT-A) correctly classifies 78.5% of the five agree testing images, outperforming similar models trained on high-quality sentiment-related hand-labeled data. This project uses textual analysis to solve the problem of training a visual sentiment classifier from a large set of multimedia data without any human annotators. This highlights the potential of unsupervised learning, which allows for greater fine-tuning, thus achieving a higher accuracy for image sentiment analysis. However, this paper over-emphasises the importance of the textual element of the input, and labels tweets only based on text, without consideration of the visual aspect.

b. Multimodal Sentiment Analysis To Explore the Structure of Emotions[2]

This multimodal sentiment analysis approach uses deep neural networks combining visual analysis and natural language processing to infer the latent emotional state of the user with a large noisy labeled dataset of Tumblr posts and a focus on prediction of the emotion word tags attached by users. This mitigates the lack of large training datasets and potential differences between users' underlying emotional states and the sentiments expressed in their posts. This approach uses Inception, a pre-trained deep CNN for image recognition, and GloVe, a regression model for unsupervised learning of word representations. A dense layer then combines the information in the two modalities and a final softmax output layer gives the probability distribution over the possible emotion word tags. This research proves that a multimodal approach combining textual and image features outperforms separate models based solely on either, with a test accuracy of 72% compared to the 40% and 69% in images-only and text-only models. This model automatically compiles dozens of english words into psychologically meaningful categories ordered by the relative frequency of the emotion being used as a tag on Tumblr, including modern slangs and phrases. However, the study is subjected to biases on social media, and is limited by expression of emotions online which may differ from how they are truly experienced.

3 The Study & Methodology

3.1 Proposed method

In this project, multimodal embedding is used in conjunction with textual analysis to provide an overall sentiment of an image. Textual representations are significant sources of information that improves sentiment prediction by providing additional semantics not necessarily found in images, seeking an explicit relationship between vision and language. There is little to no work on the method of mapping of word representations to visual representations in the form of vectors for binary sentiment analysis, and thus is novel. Furthermore, the methodology allows for unsupervised learning, which is advantageous as annotated visual representations are few and far between, compared to sizable textual word embeddings.

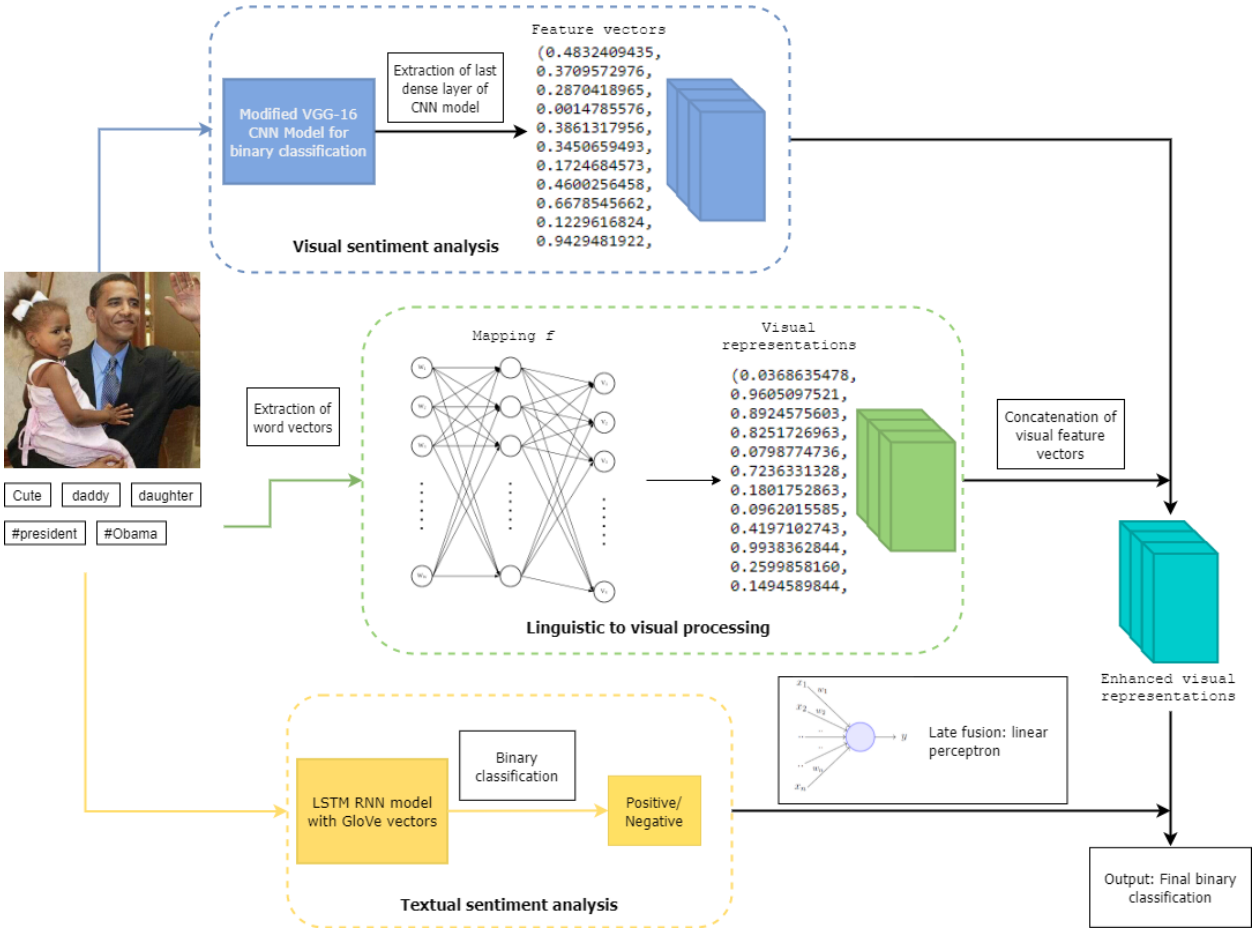

Figure 1.1

Multi-modal representation is generated through simultaneous learning; a combination of word embedding and Convolutional Neural Networks (CNNs). The mapping *f* is trained through associating the visual representation with the linguistic modality of the textual inputs from training sets; the embedded vectors from word embedding are fed into a neural network and fitted with the feature vectors from the last dense layer of the CNN.

When predicting sentiment of images, as seen in *Figure 1.1,* the mapped vectors allow for textual features to generate visual representations, which are concatenated with actual visual representations of the input image and fed into the CNN again, with a binary classification output. The textual sentiment analysis model, consisting of a Long Short-Term Memory (LSTM) architecture, generates a binary classification score (0 < 1). Both outputs will then be passed through a linear perceptron for late fusion or regression, to give a final binary classification output.

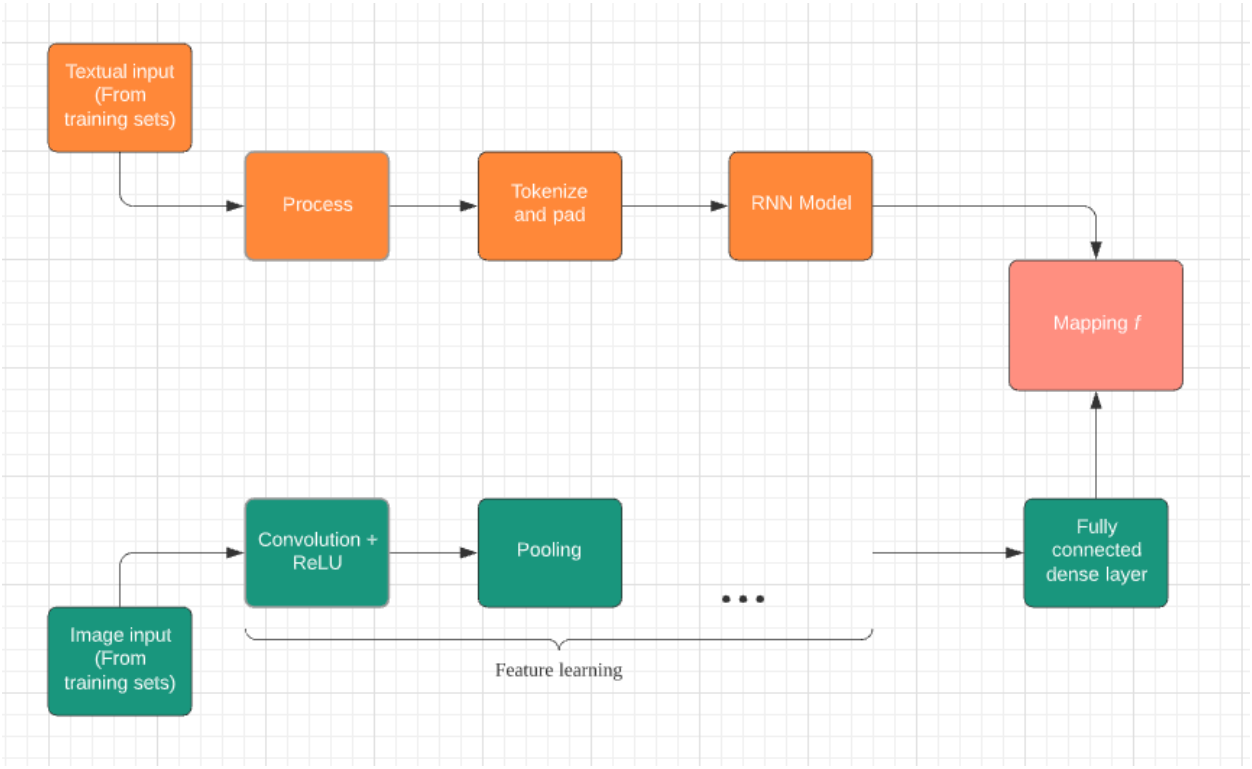

Figure 1.2 (Training)

3.2 Textual analysis

Using captions of Tweets from the Phototweet Sentiment Benchmark dataset[4], the text was processed to remove non lexical data, and then tokenized, converting words into sequences. The sequences are then padded to the max length of the sequences, to be fed into the recurrent neural network(RNN).

We used 300-dimensional GloVe vectors for the word embedding layer, which are pre-trained on the Common Crawl corpus consisting of 840B tokens and a 2.2M words vocabulary list[5]. GloVe vectors act as a non-trainable base to learn mappings of text vectors to visual feature vectors.

The RNN used is an LSTM, shown in *Figure 2*. LSTMs can process long sequences of vectors and retain relevant information to make predictions through the usage of cell states.

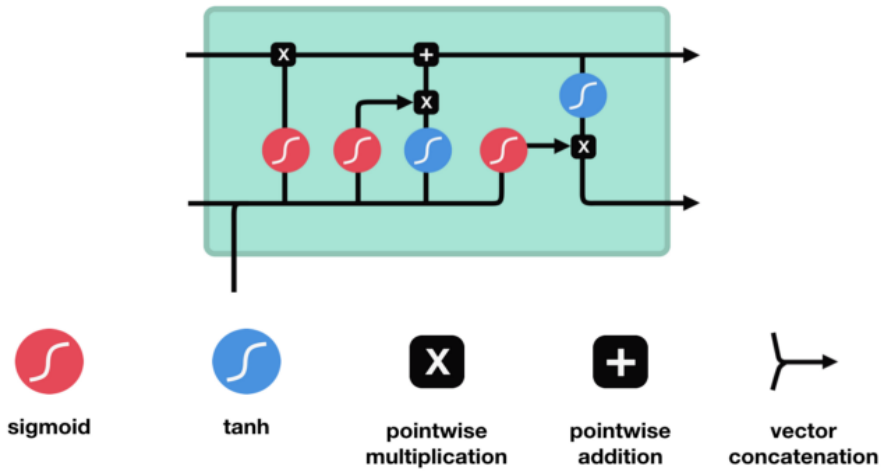

Figure 2(Phi, 2020)

3.3 Visual analysis

For feature extraction, we used a pretrained model, VGG-16 (Figure 3), which has produced stellar results especially in image classification, as the base. The last dense layer of the CNN, right before the softmax activation function, contains extracted feature vectors, which are used in **3.4**.

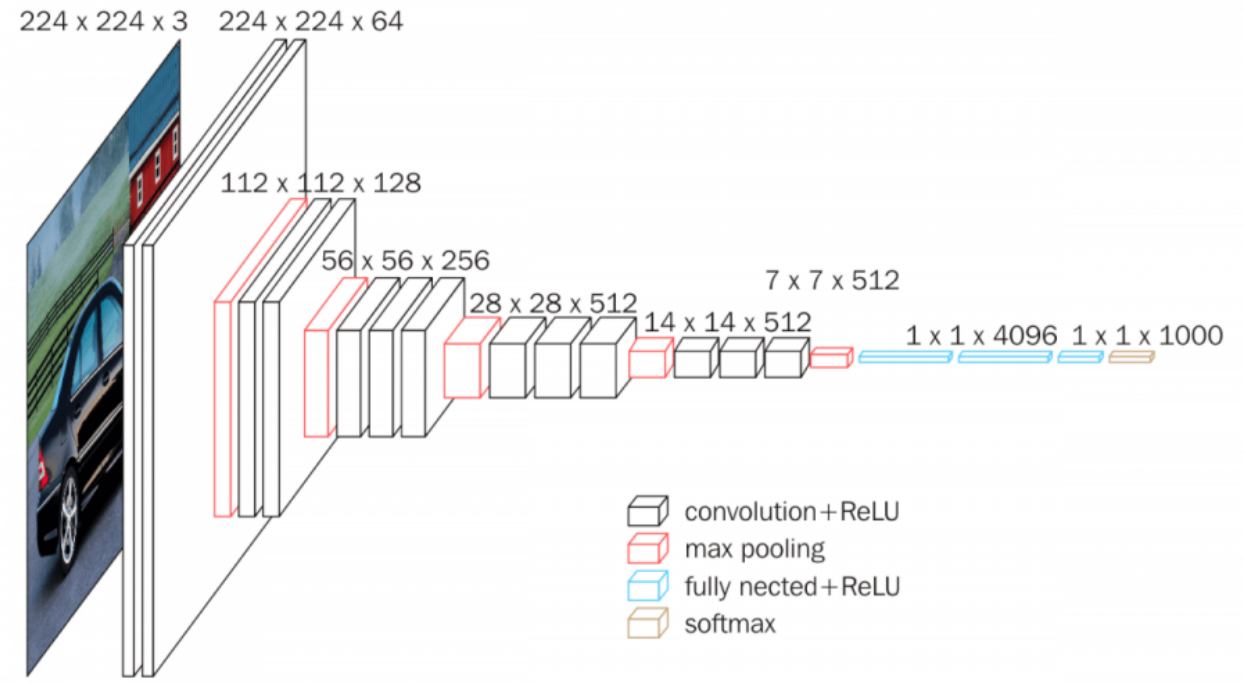

Figure 3 (Hassan, 2018)

Illustration of VGG-16 CNN model [8]

3.4 Mapping *f*

Embedded vectors from the word embedding in **3.1** were fed into a neural network. The word vectors are selected through performing a linear operation on the vectors to find the most similar vector to the sentiment of the image (negative or positive). The visual feature vector is obtained through max pooling/averaging of visual feature vectors of the corresponding sentiment. The neural network has 1 dense layer, and is fit to the vectors of the last dense layer from the CNN. In *Figure 4*, word vectors w_w, are inputted into the neural network, and outputs are of visual feature vectors v_V.

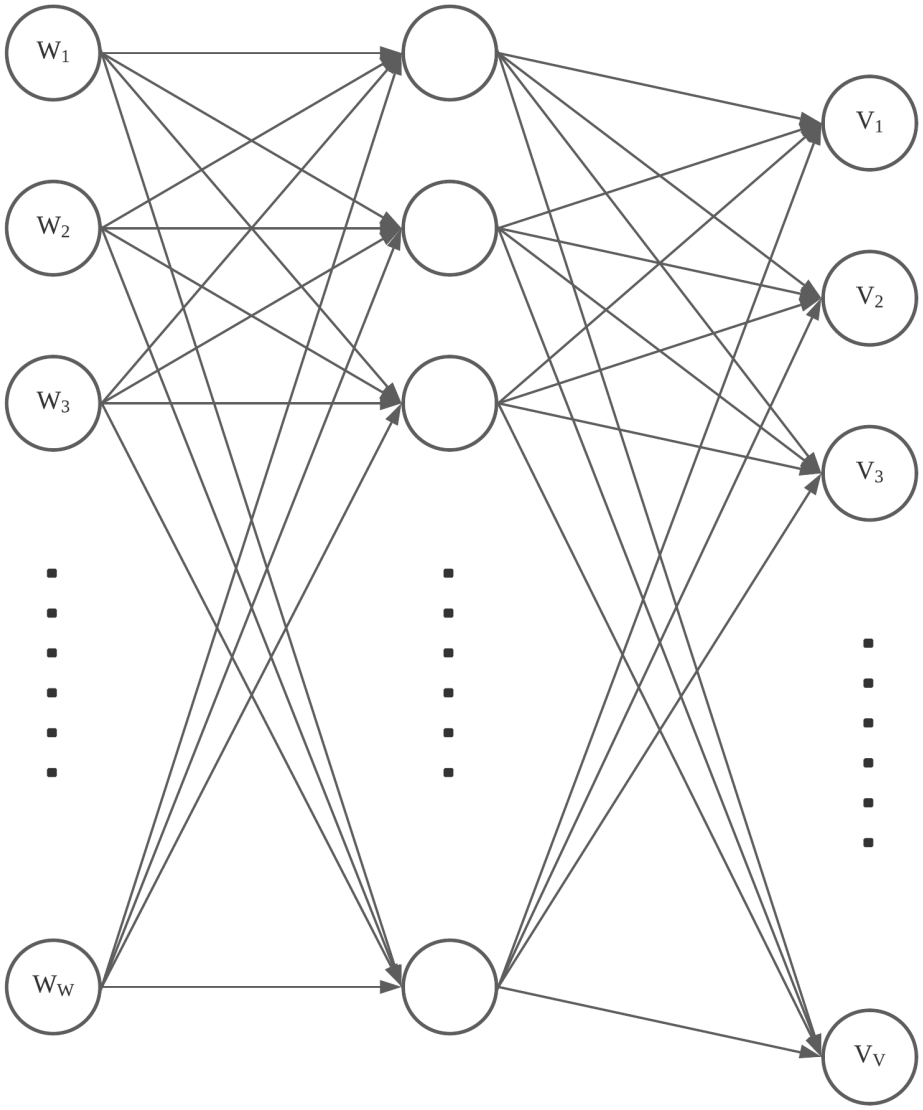

Figure 4

3.4 Dataset Preparation

We used the Phototweet Sentiment Benchmark dataset[4], which includes 603 tweets with photos and is intended for evaluating the performance of automatic sentiment prediction using features of different modalities. The tweets serve as textual input while the accompanying photos serve as visual input. *Figure 5* shows an example of a tweet.

Figure 5, image of positive sentiment with the caption:
"You should go HIV/AIDS testing!!"

4 Outcomes, Analysis & Discussions

Correctly Identified	158
Incorrectly Identified	158
Neutral (Discounted results)	287
% Accuracy (Without discounted results)	50.0
% Accuracy (With discounted results)	26.2

Table 1.1, accuracy metrics of image-only model

	Predicted: Negative	Predicted: Positive
Actual: Negative	True Negative (44 images)	False Positive (69 images)
Actual: Positive	False Negative (89 images)	True Positive (114 images)

Table 1.2, confusion matrix of image-only model (discounting neutral results)

Correctly Identified	267
Incorrectly Identified	140
Neutral (Discounted results)	196
% Accuracy (Without discounted results)	65.6
% Accuracy (With discounted results)	44.3

Table 2.1, accuracy metrics of multimodal approach

	Predicted: Negative	Predicted: Positive
Actual: Negative	True Negative (52 images)	False Positive (33 images)
Actual: Positive	False Negative (107 images)	True Positive (215 images)

Table 2.2, confusion matrix of multimodal approach (discounting neutral results)

By using the Phototweet Sentiment Benchmark[4] as test bench, we achieved an overall accuracy of 65.6% after averaging results of textual and image sentiment analysis and discounting neutral images, a 15.6% increase over the prior image-only model, demonstrating the superiority of integrated data analysis.

4.1 Challenges and limitations

During implementation of the solution, we faced 2 major challenges.

Firstly, the size of the PhotoTweet dataset is far too small, as image datasets usually have over hundreds of thousands of images. However, due to computational limits of our equipment, it was ill equipped to handle the size of bigger datasets such as the Visual Sentiment Ontology dataset, so we had to settle for the Phototweet dataset with only 603 images. This resulted in underfitting, which severely impacted the accuracy.

Secondly, extensive research had to be done beforehand, so as to gain the necessary technical expertise and experience with the various platforms and libraries. As all of us are relatively new to machine learning, grasping the requisite concepts was a hurdle that took up a considerable amount of time, such as during the implementation of VGG-16, which required many repeated tries and errors before we could get it to function.

5 Implications & Recommendations

5.1 Implications

The ability to predict overall sentiments of specific social media posts opens up a slew of opportunities for social media companies, such as tailoring post recommendations to users' moods. The general public's sentiment on specific issues can also be easily deduced, such as that on social issues like poverty, climate change and healthcare.

5.2 Recommendations

Although promising, the current Proof Of Concept (POC) approach has some flaws. Firstly, it assumes that the text has relative significance towards the image. For example, Figure 4.1 has an image of positive sentiment and a caption with no distinctive sentiments, leading to possibly inaccurate predictions with our approach, as both textual and visual sentiment analysis models use binary cross entropy loss coupled with sigmoid activations. Thus, this binary-natured classification algorithm would produce inaccurate outputs for neutral textual inputs. One possible solution would be using a softmax activation function with categorical cross entropy loss, with more labels so the model can detect spatial correlations much more accurately.

Secondly, larger training sets could be used for both models. The textual model is however a SentiStrength model, pre-trained on sample texts that exclude certain slangs customary to modern society, the inclusion of which, we believe, would greatly benefit the performance of our textual sentiment analysis. Furthermore, the image model would have far greater accuracy if trained on a larger dataset, as the current dataset causes underfitting for our model.

Figure 5, image of positive sentiment with the caption:
"Do we really need AKB48 on a poster for the Tokyo gubernatorial election?"

6 Conclusions

With research demonstrating that our late fusion multimodal POC's linear method of weighing outputs from both models works effectively, the vector-based methodology proposed in **3.1** that uses the association of semantic implications with visual features is likely to produce more accurate and reliable results in comparison. Thus, the proposed methodology is able to classify to a greater depth and expected to perform better in real world applications.

7 Bibliography

1. Vadicamo, L. V., Carrara, F. C., Cimino, A. C., Cresci, S. C., Dell'Orletta, F. D., Falchi, F. F., & Tesconi, M. T. (2017, October 1). Cross-Media Learning for Image Sentiment Analysis in the Wild. IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/8265255/
2. Hu, A. H., & Flaxman, S. F. (2018, August 19). Multimodal Sentiment Analysis To Explore the Structure of Emotions. https://sci-hub.se/10.1145/3219819.3219853
3. A. Cimino and F. Dell'Orletta. Tandem LSTM-SVM approach for sentiment analysis. In EVALITA 2016. 3, 5
4. Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel and Shih-Fu Chang. "Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs," ACM Multimedia Conference, Barcelona, Oct 2013.
5. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.
6. Kurban, R. (2020, January 18). CNN sentiment analysis. Medium. https://towardsdatascience.com/cnn-sentiment-analysis-9b1771e7cdd6
7. Phi, M. (2020, June 28). Illustrated guide TO LSTM's And GRU's: A step by step explanation. Medium. https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21
8. Hassan, M. U. (2021, February 24). VGG16 – Convolutional Network for Classification and Detection. Neurohive. https://neurohive.io/en/popular-networks/vgg16/