# MULTIMEDIA SENTIMENT ANALYSIS USING MODIFIED CNN AND RNN MODELS

By

**Youssef Saad Ghatas**

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
**MASTER OF SCIENCE**
**in**
**Computer Engineering**

FACULTY OF ENGINEERING , CAIRO UNIVERSITY
GIZA, EGYPT
2017

# MULTIMEDIA SENTIMENT ANALYSIS USING MODIFIED CNN AND RNN MODELS

By

**Youssef Saad Ghatas**

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
**MASTER OF SCIENCE**
**in**
**Computer Engineering**

Under the Supervision of

Prof. **Elsayed Hemayed**
Professor of Computer
Computer Engineering Department
Faculty of Engineering , Cairo University

FACULTY OF ENGINEERING , CAIRO UNIVERSITY
GIZA, EGYPT
2017

# MULTIMEDIA SENTIMENT ANALYSIS USING MODIFIED CNN AND RNN MODELS

By

**Youssef Saad Ghatas**

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
**MASTER OF SCIENCE**
**in**
**Computer Engineering**

Approved by the Examining Committee:

---

Prof. **Elsayed Hemayed**, Thesis Main Advisor

---

Prof. **Mohsen A. Rashwan**, Internal Examiner

---

Prof. **Reda Abd Elwahab Ahmed**, External Examiner
Professor of Computer Science, Faculty of Computers and Information, Cairo University

FACULTY OF ENGINEERING , CAIRO UNIVERSITY
GIZA, EGYPT
2017

| | | |
|---|---|---|
| **Engineer's Name:** | Youssef Saad Ghatas | |
| **Date of Birth:** | 25/06/1991 | |
| **Nationality:** | Egyptian | |
| **E-mail:** | Youssef.Ghatas@gmail.com | |
| **Phone:** | 01221198860 | |
| **Address:** | 43 Rostom Street, Helwan, Cairo | |
| **Registration Date:** | 01/10/2013 | |
| **Awarding Date:** | 2017 | |
| **Degree:** | Master of Science | |
| **Department:** | Computer Engineering | |



**Supervisors:**

Prof. Elsayed Hemayed

**Examiners:**

| | |
|---|---|
| Prof. Reda Abd Elwahab Ahmed<br>Professor of Computer Science,<br>Faculty of Computers and Information,<br>Cairo University | (External examiner) |
| Prof. Mohsen A. Rashwan | (Internal examiner) |
| Prof. Elsayed Hemayed | (Thesis main advisor) |


**Title of Thesis:**

Multimedia Sentiment Analysis using Modified CNN and RNN Models


**Key Words:**

Multimedia Sentiment Analysis; RNN; CNN; Machine Learning



**Summary:**
Multimedia Sentiment Analysis is considered a great challenge, given the diversity of data it combines and the informality challenges it produces. Firstly, this study divides the problem into textual and visual fields and study the effect of using different CNN and RNN models. Secondly, the proposed model is compared to the top candidate and the state-of-the-art model. The proposed model -which uses deep CNN model for images and RCNN model for texts- outperforms the other tested models and the state-of-the-art model on the same dataset in both accuracy and F1 score with absolute improvement of about 5% and relative error improvement of more than 25%. Finally, a sensitivity test is conducted to test the effect of different values for some important parameters of the proposed model.

# Acknowledgements

I would like to thank Dr. Elsayed Hemayed and Dr. Mohamed Aly for their support, advice and guiding throughout the thesis. I would like to thank Dina Tantawy for her continuous encouragement and help. Finally, I would like to express my sincere gratitude to Quanzeng You for providing us with a copy of his dataset.

# Dedication

This thesis is dedicated to the soul of the dearest friend Michael Wagdy. You are truly missed... never forgotten. Thanks for all your encouragement.

My thanks to my family and friends for their unconditional support, help and encouragement.

# Table of Contents


**Acknowledgements** **i**

**Dedication** **ii**

**List of Tables** **v**

**List of Figures** **vi**

**List of Symbols and Abbreviations** **viii**

**Abstract** **ix**

**1 Introduction** **1**
- 1.1 Motivation . . . 1
- 1.2 Objectives . . . 3
- 1.3 Achievements . . . 3
- 1.4 Organization of the thesis . . . 3

**2 Literature Review** **5**
- 2.1 Visual Sentiment Analysis . . . 5
  - 2.1.1 Low-level and mid-level features approaches . . . 5
  - 2.1.2 Neural network approaches . . . 5
- 2.2 Textual Sentiment Analysis . . . 7
  - 2.2.1 Lexicon-Based Approaches . . . 7
  - 2.2.2 Machine Learning Approaches . . . 9
- 2.3 Multimedia Sentiment Analysis . . . 11
- 2.4 Summary . . . 14

**3 Neural Networks** **15**
- 3.1 Convolutional Neural Networks . . . 15
  - 3.1.1 Common Layers . . . 15
  - 3.1.2 Overfitting . . . 18
  - 3.1.3 Learning process . . . 18
  - 3.1.4 Transfer Learning . . . 20
- 3.2 Recurrent Neural Networks . . . 21
  - 3.2.1 Gated Recurrent Neural Networks . . . 22
- 3.3 Vector representation of words . . . 24
  - 3.3.1 Word Embeddings Layer . . . 24
  - 3.3.2 Word2Vec Model . . . 24

**4 Proposed Model & Variations** **26**
- 4.1 Text Models . . . 26
  - 4.1.1 Embedding Layer Training Process . . . 26
  - 4.1.2 Proposed Model and Variations . . . 29

4.2 Image Models . . . 32
4.2.1 Small Model . . . 32
4.2.2 Proposed Image Model . . . 33
4.3 Combined Model . . . 34

**5 Experimental Results** **35**
5.1 Datasets . . . 35
5.1.1 Text Dataset - Tweets Dataset [TTD] . . . 35
5.1.2 Images Dataset - Flickr Dataset [FID] . . . 35
5.1.3 Combined Dataset - Image-Tweets Combined Dataset [TCD] . . . 37
5.2 Models Experiments & Comparisons . . . 39
5.2.1 Evaluation . . . 39
5.2.2 Implementation Details . . . 39
5.2.3 Text models experiments . . . 40
5.2.4 Image models experiments . . . 45
5.2.5 Combined models experiments . . . 50
5.3 Proposed Model Tests . . . 52
5.3.1 Early Fusion Test . . . 52
5.3.2 Text Model Sensitivity Tests . . . 53
5.4 Summarized Comparison between Single and combined models . . . 57
5.5 Results on samples and model limitations . . . 59

**6 Conclusion and Future Work** **61**
6.1 Conclusion . . . 61
6.2 Future work . . . 61

**References** **63**

**A Saving Intermediate Features** **68**
A.1 Used Machines . . . 68
A.2 Methodology . . . 68
A.3 Comparison and Analysis . . . 69
A.3.1 Notes and Limitations . . . 70

# List of Tables


2.1 Results of the models proposed by You et al. [51] . . . . . . . . . . . . . . 14

3.1 Transfer Learning - Common Scenarios . . . . . . . . . . . . . . . . . . . . 20

4.1 Text preprocessing - Example . . . . . . . . . . . . . . . . . . . . . . . . . 28

5.1 TTD - Samples . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35
5.2 Confusion Matrix . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 39
5.3 Comparison between text models . . . . . . . . . . . . . . . . . . . . . . . 45
5.4 Comparison between image models . . . . . . . . . . . . . . . . . . . . . 49
5.5 Comparison between combined models . . . . . . . . . . . . . . . . . . . 51
5.6 Comparison between Early Fusion Model and the Proposed Model . . . . 53
5.7 Sensitivity test - Removing dropout layers - Results . . . . . . . . . . . . 54
5.8 Sensitivity test - Convolutional layer (filter size) - Results . . . . . . . . . 55
5.9 Sensitivity test - Convolutional layer (number of filters) - Results . . . . . 55
5.10 Sensitivity test - Embedding Layer - Results . . . . . . . . . . . . . . . . 56
5.11 Comparison between multiple single and combined models . . . . . . . . 58

A.1 Machines' specifications . . . . . . . . . . . . . . . . . . . . . . . . . . . 68
A.2 Comparison between loading the complete models and saving intermediate outputs . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 69

# List of Figures


1.1 Ambiguous text . . . 2
1.2 Ambiguous image . . . 2

2.1 Model Architecture and methodology used by Xu et al. [47] . . . 6
2.2 Model Architecture used by You et al. [50] . . . 7
2.3 Model Architecture used by Campos et al. [5] . . . 8
2.4 Model architecture with two channels for an example sentence used by Kim [22] . . . 11
2.5 Model Architecture used by Stojanovski et al. [42] . . . 12
2.6 Model Architecture used by Yu et al. [52] . . . 13
2.7 Model Architecture used by You et al. [51] . . . 13
2.8 Model architecture used by Le and Mikolov [26] . . . 13

3.1 Feedforward Neural Network . . . 15
3.2 Pooling Layer . . . 16
3.3 Fully connected layer . . . 16
3.4 Recurrent Neural Network - Loop Form (from [33]) . . . 21
3.5 Recurrent Neural Network - Sequence Form (from [33]) . . . 21
3.6 Simple RNN - using simple tanh function (from [33]) . . . 23
3.7 LSTM - using complex structure (from [33]) . . . 23
3.8 Word2Vec diagram . . . 24

4.1 Combined model architecture with sample inputs . . . 26
4.2 Preprocessing, Word2Vec Model and Embedding Layer . . . 27
4.3 NN Model - Simple NN Model . . . 29
4.4 CNN Model - Concatenated Multiple CNN Model . . . 30
4.5 RNN Model 1 - Simple RNN (LSTM) . . . 30
4.6 RNN Model 2 - Simple RNN (GRU) . . . 30
4.7 RNN Model 3 - CRNN . . . 31
4.8 Proposed Text Model - RCNN . . . 32
4.9 Small Model - Image Simple CNN Model . . . 32
4.10 Pretrained model used - Cut & Modified dimensions . . . 33
4.11 Proposed image model - CNN Top Model - Using a pretrained model . . . 34
4.12 Combined model . . . 34

5.1 Flickr Dataset - Negative Samples . . . 36
5.2 Flickr Dataset - Positive Samples . . . 36
5.3 Positive samples from TCD . . . 37
5.4 Negative samples from TCD . . . 38
5.5 NN Model - Learning graph . . . 40
5.6 CNN Model - Learning graph . . . 41
5.7 RNN Model 1 - Learning graph . . . 42
5.8 RNN Model 2 - Learning graph . . . 42
5.9 RNN Model 3 - Learning graph . . . 43

5.10 RNN Model 4 - Learning graph . . . . . . . . . . . . . . . . . . . . . . . . . 43
5.11 Proposed Text Model - Learning graph . . . . . . . . . . . . . . . . . . . . . . 44
5.12 Small Model - Learning graph - First Phase . . . . . . . . . . . . . . . . . . 46
5.13 Small Model - Learning graph - Second Phase . . . . . . . . . . . . . . . . . 47
5.14 Small Model - Learning graph - Third Phase . . . . . . . . . . . . . . . . . . 47
5.15 Proposed Image Model - Learning graphs . . . . . . . . . . . . . . . . . . . 48
5.16 Combined models - learning graphs . . . . . . . . . . . . . . . . . . . . . . 51
5.17 Early Fusion - Model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 52
5.18 Early Fusion - learning graphs . . . . . . . . . . . . . . . . . . . . . . . . . 52
5.19 Proposed Text Model - RCNN . . . . . . . . . . . . . . . . . . . . . . . . . 53
5.20 Sensitivity test - Removing dropout layers - Learning graph . . . . . . . . 53
5.21 Sensitivity test - Removing dropout layers - Results . . . . . . . . . . . . . 54
5.22 Sensitivity test - Convolutional layer (filter size) - Results . . . . . . . . . 55
5.23 Sensitivity test - Convolutional layer (number of filters) - Results . . . . . 56
5.24 Sensitivity test - Embedding Layer - Results . . . . . . . . . . . . . . . . . 57
5.25 Some top confident samples . . . . . . . . . . . . . . . . . . . . . . . . . . 59
5.26 Misclassified samples . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 60

A.1 Combined model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 68

# List of Symbols and Abbreviations

| | |
|---|---|
| ANP | Adjective Noun Pair |
| BoW | Bag of Words |
| CCR | Cross-modality Consistent Regression |
| CNN | Convolutional Neural Network |
| CRNN | Convolutional-Recurrent Neural Network |
| FC | Fully Connected Layer. |
| FID | Flickr Image Dataset |
| FN | False Negative |
| FP | False Positive |
| GRU | Gated Recurrent Unit |
| HSV | Hue, Saturation, Value |
| LSTM | Long Short Term Memory |
| NLP | Natural Language Processing |
| NN | Neural Network |
| PReLU | Parametric Rectified Linear Unit |
| RCNN | Recurrent-Convolutional Neural Network |
| ReLU | Rectified Linear Unit |
| RNN | Recurrent Neural Network |
| SVM | Support Vector Machine |
| TCD | Twitter Combined Dataset |
| TN | True Negative |
| TP | True Positive |
| TTD | Twitter Text Dataset |

# Abstract

Retrieving the emotional effect of images or texts is very important to be able to handle the published content on the internet. It can be used for product selection, politics and public opinion statistics.

In spite that a lot of efforts have been made to address textual and visual ordinary classification problems, much less efforts have targeted multimedia sentiment analysis. Traditional studies of sentiment analysis targeted either images or texts, yet, it is inadequate in the era of social multimedia. It doesn't only limit the scope of the targeted content, but it also neglects the way our human brain works to combine different data. Multimedia sentiment analysis is considered a challenging task due to its high level of classification, in addition to the informalities from which the social media content suffers.

In this study, we address this problem on Twitter social multimedia content that contains both textual and visual data. We present an end-to-end complete neural network model to solve the problem of multimedia sentiment analysis. Firstly, we introduce a deep Convolutional Neural Network (CNN) for visual sentiment analysis. We compare it to the state-of-the-art model for sentiment analysis of Twitter's images. Secondly, a modified Recurrent Neural Network (RCNN) -which uses both convolutional and recurrent layers- is proposed for textual sentiment analysis. It is compared to multiple variations to test the effect of different architectures on the model. In these tests, we test two recurrent layers, namely, Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) and we test adding a convolutional layer before and after the recurrent layer.

To handle the mixed inputs of images and texts we test early-fusion and late-fusion between the two proposed models and we also test the effect of fine-tuning.

To test the effect of different values for the parameters, we conduct a sensitivity test for the proposed text model which shows the robustness of the proposed model. It also shows that the chosen parameters are enough for the models' capacity.

Finally, we analyze the proposed model using different test samples to show its points of strength and limitations. In comparison to the state-of-the-art model, the proposed model shows superior performance with absolute improvement of 5% and relative error improvement of 25% on the same test set.

# Chapter 1: Introduction

Online social networks have gained a lot of importance in our daily life. Twitter, Facebook and other social media channels have become the major sources of information, news and opinions. Most users find it a secure way to express themselves freely. Over the time, it became an expressive and unbiased way to extract opinions about different topics. It has a wide variety of applications. For example:

1. Product Selection based on the reviews of customers: the reviews of customers are used to estimate the acceptance or refusal of a product [10].

2. Predicting the results of elections: which can be used by different parties to modify their vision according to the opinions of the people [4].

## 1.1 Motivation

Retrieving sentiment analysis for social media content is considered a difficult task because it is a high level classification problem. In addition to that:

1. The same input -whether an image or a text- can have multiple meanings given different environments and events.

2. Social media data suffers from informality.

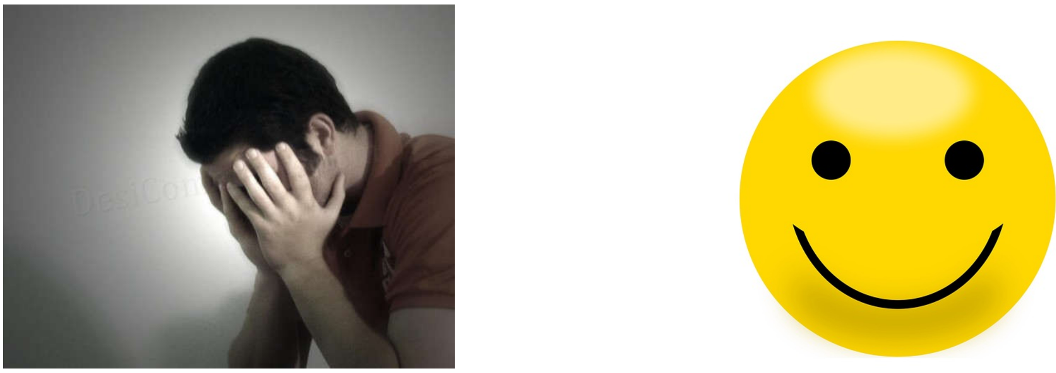
3. Images may be very different in size and quality.

4. Texts may contain non-traditional English words and abbreviations.

Textual sentiment analysis alone has become inadequate due to the variety of forms that social networks provide for users to express themselves. Accompanied with other forms of media such as images, it's more likely to be able to express and convey people's subtle feelings. For example, two extreme examples are illustrated to show the importance of using both text and image data. In figures 1.1a and 1.1b, text alone is useless, however, it's very easy to get the sentiment using the images. In contrast, in figures 1.2a and 1.2b, sentiment depends mainly on the text as it's more illustrative.

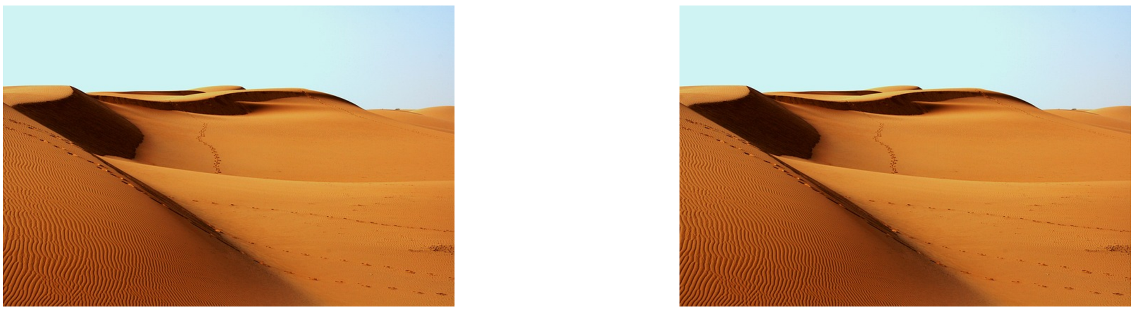
Convolutional Neural Networks (CNNs) have dominated a lot of image classification and recognition problems. They achieve state-of-the-art results in an increasing number of computer vision tasks including: scene classification [8], face recognition [1] and face detection [28].

Recurrent Neural Networks (RNNs) are appropriate for sequence problems, where the output relies on some sequence(s) of inputs. Instead of dealing with fixed-size inputs to get fixed-size outputs, they deal with variable size of sequences and relate it to one or more outputs. That's why RNNs have been used thoroughly in solving the problems of sequences including text classification [25], text completion [43], question answering[16] and speech recognition [13].



(a) “My feeling right now...” (b) “My feeling right now...”

Figure 1.1: Ambiguous text



(a) “Feeling lost :(” (b) “A great journey to the desert, loved it <3”

Figure 1.2: Ambiguous image