arXiv:1808.07235v1 [cs.CL] 22 Aug 2018

Finding Good Representations of Emotions for Text Classification

by

Ji Ho Park

A Thesis Submitted to
The Hong Kong University of Science and Technology
in Partial Fulfillment of the Requirements for
the Degree of Master of Philosophy
in the Department of Electronic and Computer Engineering

August 2018, Hong Kong

Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Ji Ho Park

August 2018

Finding Good Representations of Emotions for Text Classification

by

Ji Ho Park

This is to certify that I have examined the above MPhil thesis

and have found that it is complete and satisfactory in all respects,

and that any and all revisions required by

the thesis examination committee have been made.

Prof. Pascale Fung, Thesis Supervisor

Prof. Bert Shi, Head of Department

Thesis Examination Committee

1. Prof. Bert Shi (Chairperson)	Department of Electronic and Computer Engineering
2. Prof. Pascale Fung (Supervisor)	Department of Electronic and Computer Engineering
3. Prof. Chiew Lan Tai	Department of Computer Science and Engineering

Department of Electronic and Computer Engineering

August 2018

ACKNOWLEDGMENTS

Firstly I would like to thank my advisor Professor Pascale Fung for leading me into this field and guide me throughout my research. These two years have been a life-changing experience. I have never learned so much in such a short period of time and appreciate all the effort by her to provide me with the necessary support and advice. I am proud to say that I was advised by Pascale and passed her criteria for a research thesis.

I appreciate Professor Bert Shi and Professor Chiew-Lan Tai for taking their times to be in the thesis supervision committee. I hope you enjoy reading my thesis and following my defense presentation. I would also like to thank Professor Yangqiu Song for giving me a good feedback and advice on my research.

I thank my friend Anik Dey for introducing me to the lab and for being a good mentor and friend throughout my time in HKUST. Also, without the collaboration of Xu Peng, Jay Shin, and Dario Bertero, my thesis would not have been completed. I appreciate all their times discussing and working with me on these works. I feel grateful for all the former and current members of our lab who always support me and openly discussed each other's research. I hope this kind of culture with continue to exist and wish the best in their research as well.

I would like to send my love to parents and brother who are mostly away from me but always give me unconditional support. All the trust and love given by them have been the most important thing in my life. I also want to express my gratitude to my girlfriend Tina Chim because without her next to me for the past 1.5 years it would have been impossible for me to finish my research this well.

Lastly, I want to acknowledge myself for being so persevering and proactive. There were some hard times but I eventually overcame them and achieved many things during the two years. I hope the future self can read these words and be reminded that nothing can beat consistent daily efforts. I hope this thesis is not the end but a start of many more exciting experiences in my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

FINDING GOOD REPRESENTATIONS OF EMOTIONS FOR TEXT CLASSIFICATION

by

JI HO PARK

Department of Electronic and Computer Engineering

The Hong Kong University of Science and Technology

ABSTRACT

It is important for machines to interpret human emotions properly for better human-machine communications, as emotion is an essential part of human-to-human communications. One aspect of emotion is reflected in the language we use. How to represent emotions in texts is a challenge in natural language processing (NLP). Although continuous vector representations like word2vec have become the new norm for NLP problems, their limitations are that they do not take emotions into consideration and can unintentionally contain bias toward certain identities like different genders.

This thesis focuses on improving existing representations in both word and sentence levels by explicitly taking emotions inside text and model bias into account in their training process. Our improved representations can help to build more robust machine learning models for affect-related text classification like sentiment/emotion analysis and abusive language detection.

We first propose representations called emotional word vectors (EVEC), which is learned from a convolutional neural network model with an emotion-labeled corpus, which is constructed using hashtags. Secondly, we extend to learning sentence-level representa-

tions by training a bidirectional Long Short-Term Memory model with a huge corpus of texts with the pseudo task of recognizing emojis. We evaluate both representations by performing both qualitative and quantitative analysis and also report high-ranked results in the Semantic Evaluation (SemEval2018) competition. Our results show that, with the representations trained from millions of tweets with weakly supervised labels such as hashtags and emojis, we can solve sentiment/emotion analysis tasks more effectively.

Lastly, as examples of model bias in representations of existing approaches, we explore a specific problem of automatic detection of abusive language (also known as hate speech). We address the issue of gender bias in various neural network models by conducting experiments to measure and reduce those biases in the representations in order to build more robust classification models.

CHAPTER 1

INTRODUCTION

Emotions play an important role in our daily communications. Humans have naturally evolved to express and perceive them in different ways, such as facial expressions, tones of voice, and choice of words. For this reason, developing a sense of empathy toward other people is an essential skill for communicating effectively. Emotions inside language increase the complexity since they not only depend on the semantics but also are inherently subjective, ambiguous, and implicit. For example, the sentence, *"I have not eaten alone for three days,"* merely state a fact that the person consumed food by themselves for a period of time. However, naturally, we can imagine the emotion of the speaker and say, *"oh that must have been pretty lonely,"* and then ask that person to eat together next time. This is called being empathetic, able to understand what the others are feeling and how to correspond to that in a conversation.

Despite the difficulty, accounting for emotions is important in building a machine that truly understands natural language, especially for tasks that are directly related to affect recognition such as sentiment/emotion analysis and abusive language detection, and also those involving human-computer interactions such as dialogue systems and chatbots [22]. As humans can naturally capture and express different emotions in texts, machines should be able to learn how to infer them as well.

Some people may think why the world needs empathetic machines that understand human emotions. Popular NLP topics like task-oriented dialogue systems or question and answering, do not seem to be directly related to emotions. However, we believe that machines will take a more active, closer role in supporting humans in the future. Personal assistants like Siri will develop to learn how to make more complex conversations and the expectation of users may grow higher.

To give a more concrete example, many people in healthcare are trying to develop robots that will assist elderly people. This is inevitable due to the growing population of elderly people because of the advancement of medicine and hospitals. Those robots may not only take care of their physical abilities but also their mental states by being a friend and making a conversation.

When training NLP models, such as chatbots, things do not always go as intended. Famous incident of Microsoft chatbot Tay, which learned directly from users' tweets without any filtering and started becoming racist and spitting out abusive language, gave us a good lesson that a lot of conversation data is not all we need. We also need to be aware of what the machines are learning in terms of empathy. As an extension of this, teaching emotions to machines can include social values and ethics. Understanding what is acceptable to say in the society or when you should be angry.

Representations are the first step to teach machines to understand how humans see the world. In other words, representations are ways of expressing the raw input data in a way that machine learning models can effectively deal with. For example, colored images are converted into two-dimensional matrices with RGB pixel values, because that is how our eye's retina perceive the world visually. Good representations should contain essential information of the data and be a useful input for statistical models to solve problems like classification and regression.

Finding good representations of texts is very challenging since texts are sequences of words which are represented in a discrete space of the vocabulary. The most naive way is to create a dictionary and treat each word as a separate feature inside a sentence. This approach, called the bag-of-words, surprisingly works pretty well for some basic tasks, but do not scale to more complicated natural language processing (NLP) tasks, since it takes away the word order and does not learn much about the relationship among words.

Many past works have investigated in finding the mapping of words [39, 54] or sentences [31] to continuous spaces so that each text can be represented by a fixed-size, real-valued vector. Using these vector representations of texts are more effective compared to traditional linguistic features such as word/char n-grams since these vectors have much lower dimensions and encode richer syntactic and semantic information of texts.

Nevertheless, work focusing on learning representations of emotions inside texts has not been thoroughly explored yet. Previous works mentioned above mostly considered syntax and semantics, which may not be enough for affect-related tasks. Recently, a few works [20, 69] started to show that including sentiment and emotional information in learning representations can be very useful for many relevant tasks. Our work is highly related to these efforts and more literature review will be introduced in Chapter 2.

Without a huge corpus, learning robust representations with a powerful generalizable ability is difficult, since language is inherently very diverse, with endless ways to express one's intention and emotion. However, thanks to the endless stream of social media such as Twitter and Facebook, researchers nowadays are lucky enough to have access to almost an unlimited number of texts generated every day. However, annotating these texts with emotion or sentiment human labels is very expensive and difficult. For this reason, a lot of work naturally focused on finding direct or indirect evidence of emotion inside each text, such as hashtags and emoticons [67, 73], and found them useful to distantly label an emotion of each text. Furthermore, the recent popular culture of using emojis inside social media posts and messages provide us even richer evidence of diverse emotions [20, 78]. Our work also makes good use of these methods to utilize the streams of data for learning good representations of emotions inside texts.

To learn a good representation from a huge corpus, An appropriate statistical model with sufficient learning capability should be selected. Recently, various deep learning models have been proven to be very powerful for representation learning in the area of natural language, speech recognition, vision, etc. [4], especially with a huge amount of training data. We explore various deep learning models, such as convolutional neural networks (CNN) and Long Short-Term Memory (LSTM) networks, to learn good representations of emotions of texts. We propose methods for both word and sentence levels since we assume that different levels of representation can capture different information. To prove the effectiveness of these representations, we propose methods to apply them to other relevant text classification problems such as sentiment/emotion analysis and present the performance of our representations. The dataset we use for evaluation includes excerpts of interviews, social media posts (tweets), and online comments.

Additionally, we broaden the thesis scope by addressing a more specific application, automatic abusive language detection. Abusive language, caused by negative emotions such as anger, fear, and hatred, is an important social issue directly related to our lives. As the number of posts generated on the Internet every day significantly exceeds the capability of human moderators, automatic abusive language detection has become a major demand for many companies such as Google and Facebook. In our work, we apply methods to automatically learn from different levels of representations like characters and words to classify abusive language. Moreover, we discuss the problem of gender bias in the representations of various neural networks learned from existing abusive language datasets and explore ways to reduce those bias to improve the robustness of the representations captured by those models.

The rest of the thesis is organized as follows:

- Chapter 2 (Background) provides important reviews in both psychology and natural language processing literature that are fundamentals to our work.

- Chapter 3 (Emotion Representations in Words) introduces emotional word vectors learned from a hashtag corpus and compare them with other widely used word vectors.

- Chapter 4 (Emotion Representations in Sentences) presents a method of learning good sentence representations from an emoji cluster corpus and show their effectiveness in sentiment/emotion analysis.

- Chapter 5 (Abusive Language Detection) discusses our approach of solving automatic abusive language detection from existing public datasets and address the issue of gender bias in the representations of various classification models.

- Chapter 6 (Conclusion) summarizes the results and the significance of learning good representations of emotion in many text classification tasks.

CHAPTER 2

BACKGROUND

Before proposing our methods, we present some important background knowledge that is fundamental to our work. First of all, we provide some literature review on representations of words and sentences in natural language processing (NLP) research. As mentioned in the introduction, representations are the first things to consider when building a machine learning model, since they fundamentally change how these models can recognize the given input data. We review the existing representation approaches, discuss what are the limitations of them, and connect their relevance to sentiment and emotion.

2.1 Categorical Representations

Categorical representations are the most simple way to represent texts. One-hot encodings represent each word in the vocabulary as a binary variable. In a V-dimensional vector (V is the size of the vocabulary), the index of the corresponding word is marked with an integer value 1. All other columns are marked with 0 (Figure 2.1). Bag-of-word representation is an extension of one-hot. It simply sums up the one-hot representations of words in the sentence. Categorical representations are simple and intuitive, but their limitations are that they do not consider any relational information among words and ignore word orders inside a sentence. Also, in the English language there exists a huge number of words. For this reason, naively representing a word inside a huge vocabulary, such as a one-hot vector of 100,000 words, will easily suffer from the curse of dimensionality.

'Hello' = [1, 0, 0]
'World' = [0, 1, 0]
"Hello World" = [1, 1, 0]

Figure 2.1. Illustration of 'Hello' and 'World' as one-hot and "Hello World" as bag-of–word. Imagine a 100,000 dimensional vector for a large sized vocabulary.