作者简介及博士学位论文中英文摘要

论文题目：基于表示学习的文本情感分析研究

作者简介：唐都钰，男，1988 年 10 月出生，2012 年 9 月师从于哈尔滨工业大学秦兵教授，于 2016 年 7 月获博士学位。

中 文 摘 要

文本情感分析是自然语言处理领域的热点问题之一，目的在于自动地从文本中发现和归纳主观情感信息。情感分析的一个核心问题是如何有效地表示各层次(如词语、句子、篇章)文本的情感语义，其重要性主要体现为两个方面。一方面，理解文本的情感需要让机器理解人类语言中所表达的语义，文本的语义表示是人类和机器之间信息传递的桥梁；另一方面，目前基于机器学习的方法在多个层次的情感分析任务上取得了非常出色的效果，而一个机器学习算法的性能很大程度上取决于特征表示的好坏，因此文本的表示直接影响情感分析的性能。鉴于此，本课题在文本情感分析的背景下研究各层次文本的表示，以提升各层次情感分析的性能。本课题充分考虑文本的结构性和语义组合性，将研究内容概括为以下四个方面：

1. 面向情感分析的词向量学习。传统的词向量学习算法通常基于分布假设，认为一个词语的语义是由它的上下文词语决定。相应地，情感类别相反但上下文词语相近的词语(如“好”和“坏”)容易被映射到相近的词向量，这会直接影响情感分析的性能。为了解决这个问题，本课题学习面向情感分析的词向量，在学习词向量的过程中同时利用词语的上下文信息和文本的情感类别信息，使形如“好”和“坏”等情感极性相反的词语在语义空间中被区分开。

2. 基于深层记忆网络的对象级情感分析。上下文词语对判断一个对象(aspect)的情感极性至关重要，而不同的上下文词语对于对象的重要程度是不同的。例如在句子“食物好吃但是服务很糟糕”中，对于对象“食物”,“好吃”的重要程度大于“糟糕”；而对于对象“服务”,“糟糕”的重要程度要大于“好吃”。基于上述分析，我们提出了基于深层记忆网络的对象级情感分类模型，该模型首先学习得到每个上下文词语的重要性，随后利用该重要性信息计算对象在当前句子环境下的的表示，并判断对象的情感极性。

3. 基于切分-分类联合模型的句子级情感分析。很多的情感表达是包含多个词语的复合词组，如情感否定(如“不喜欢”)、情感加强(如“非常好”)和非语义组合现象(如“等不及(cannot wait to)”)。为了处理包含多个词语的复合情感表达，我们提出了切分-分类联合模型。该联合模型在仅有句子级情感标签的条件下，自动学习句子内部各词语的切分情况，并利用切分结果判断句子的情感极性。

4. 融合用户和产品偏好的篇章级情感分析。分析篇章的整体情感是情感分析领域最早被研究的问题之一。大多数的已有工作仅使用篇章的文本信息，而忽略了用户和产品的偏好对文本情感极性/强度的影响。我们认为用户和产品的打分偏好和用词偏好对文本情感分析十分重要，并在大规模数据集上验证了这两种偏好。为了在篇章级情感分析中有效利用用户和产品的偏好信息，我们把用户和产品映射到语义向量空间，并融入基于神经网络的语义组合框架中进行情感分析。

综上，本课题在文本情感分析的背景下，从词语、对象、句子和篇章多个角度研究文本的语义表示，以提高情感分析的性能。希望本研究能够对文本情感分析和自然语言处理领域的学者提供一定的参考作用。

关键词：文本情感分析；表示学习；面向文本情感分析的词向量；深层记忆网络；切分分类联合模型

Representation Learning For Sentiment Analysis

TANG Duyu

ABSTRACT

Sentiment analysis is a fascinating task in natural language processing, aiming at extracting and organizing sentiment information from text in an automatic way. A fundamental problem in sentiment analysis is how to effectively represent the semantic meaning of texts of different granularity like word, sentence and document. The importance of learning semantic meaning of texts could be reflected from two aspects. First, teaching a machine learner to determine the sentiment of text requires a deep understanding of the semantics of texts. Semantic representation of texts could be viewed as the bridge to communicate between computer and human being. Second, machine learning approaches have achieved promising performances on a variety of sentiment analysis tasks. As the performance of a machine learner heavily depends on the choice of data representation, it is desirable to learn powerful text feature for sentiment analysis tasks. In this thesis we focus on understanding the semantics of texts and developing powerful representation learning methods to improve the performance of sentiment analysis. Based on the structural and compositional properties of natural language, we summarize our research works into the following four aspects in this thesis.

1. Sentiment-specific word embedding (SSWE). Traditional word embedding learning methods are typically based on the distributional hypothesis, which represents the meaning of a word with its surrounding context words. Accordingly, the words with similar context words but different sentiment labels like “good” and “bad” will be mapped into close vectors in the embedding space. This is problematic for sentiment analysis. To address this problem, we propose to learning SSWE, which simultaneously captures contexts of words and sentiment polarity of sentences. As a result, the words like , so that“good” and “bad” could be separated to opposite ends of the spectrum.

2. Aspect level sentiment classification with deep memory network. Context words are very important for determining the sentiment polarity of an aspect. We believe that context words do not contribute equally with regard to an aspect. For instance, in sentence “great food but the service was dreadful!”, “great” is a more important clue than“dreadful” for the aspect “food”. On the contrary, “dreadful” is more important than“great” for the aspect “service”. To capture this information, we develop a deep memory network, which first learns the importance/weight of each context word and then utilizes this information to calculate continuous text representation.

3. A joint segmentation and classification framework for sentence level sentiment classification. Negation and intensification are fundamental phenomenons in sentiment analysis. Many sentiment expressions include multiple words like “not good” and “very happy”. To handle this, we introduce a joint segmentation and classification framework, which could first segment a sentence to a list of computational units and them use the result for sentiment classification without using any segmentation annotations.

4. Document level sentiment classification with user and product information. Existing approaches typically only use text information while ignoring the important influences of users and products. We suggest that the bias of user and product is helpful for sentiment analysis, and validate the influences of users and products in terms of sentiment and text on massive reviews. We model users and products with vector space models, and integrate them into a neural network framework for sentiment analysis.

In summary, we learn semantic representations of words, aspects, sentences and document to improve the performance of sentiment analysis. We hope that our research could be helpful to the researcher in the area of sentiment analysis.

Key words: sentiment analysis, representation learning, sentiment-specific word embedding, deep memory network, joint segmentation and classification