国内外现状

现在的用户分类主要由**机器学习**所主导。研究方面主要分为两部分。**基于文本特征的性别分类**，用户性别分类已经在几个领域进行了广泛研究，如博客（Peersmanet al。，2011;Gianfortoni等，2011）[16]，电子邮件（Mohanmad等，2011），而YouTube（Filippova，2012）[15]和微博（Liu et al。，2013）。最近，一些研究集中在一些具体的性别分类应用场景，如多语言性别分类（Ciot等，2013; Alowibdi等，2013）[14][1]，推断性别的人群（Nguyen等，2014） 和互动性别分类（Li等，2015）[24]。**基于文本特征的年龄分类**，用户年龄分类已经在两个主要领域中被研究，即博客（Burger和Hender son，2006）[8]和社交媒体（Machinnon和Warren，2006）[26]。在博客领域，Schler et al。 （2006）关注从博客文本提取的文本特征，例如词语特征和POS风格特征。 Burger和Henderson（2006）[8]探索了一些社交功能，例如与博客时代相关的位置，时间和朋友特征。其他研究，如Rosenthal和McKeown（2011）[36]和Goswamiet al。 （2009）探讨自动年龄分类中的文本和社会特征。在社交媒体领域，Mackinnon和Warren（2006）[26]探索了一些社会特征，即用户之间的关系以预测用户的年龄和社交网络中的居住国。Peersman et al。 （2011）将文本分类方法应用于具有文本特征的年龄分类只有，即单字和双字。最近，Marquardt等人（2014）[28]提出了多标签分类方法来预测文本中作者的性别和年龄。具体来说，除此之外词的特征，他们也采取一些情感和情感特征在他们的方法。

一些其他用户属性，如用户位置（Cheng等人，2010），政治取向（Rao等人，2010）和用户职业类别预测（Preotiuc-Pietro等人，2015）也近年来被广泛研究。

但是几乎所有的之前的研究都存在一个问题，没有对用户生成的文本的风格进行区分。例如：社交媒体中的主页包含至少四种文本，即原始消息，转推消息，来自其他人的评论和对其他人的评论。如果不对文本的风格进行区分，这可能会影响分类性能。

在本文中，我们通过提出一种称为集合LSTM循环神经网络的新方法来解决上述挑战。具体来说，我们首先将每种文本样式的特征作为单独的文本视角。然后，我们分别为每个文本视角训练长短期存储器（LSTM）网络。 第三，我们添加一个合并层，通过联合学习来合并所有的LSTM表示，以融合所有的文本知识。实证研究表明我们的方法表现远优于许多强基线方法。

**用户分类算法比较**

单文本视角

**ME**: 最大熵分类器与所有参数默认。

**CNN**: 基本的bow-CNN提出 (Johnson and Zhang, 2014)。

**Parallel CNN**: bow-CNN的扩展，其具有并行的两个或更多个卷积层以学习多种类型的小文本区域的嵌入(Johnson and Zhang,2014)。

**LSTM**: 单文本视角 LSTM

多文本视角

**RandomForest**: (Strobl el al 2007)提出的流行的集成学习方法。在我们的实现中，我们训练多个决策树分类器并使用随机森林算法来组合它们。

**Adaboost**: 一种流行的集合学习方法（Zhu et al。，2009）。在我们的实现中，我们混合所有文本视角的数据并使用每个词特征形成弱分类器，然后将所有特征分类器与adaboost算法组合。

**Voting LSTM**: 我们首先使用每个单一文本视角来训练LSTM分类器，然后使用投票规则（Kuncheva和Rodriguez，2014）将所有单视角LSTM分类器的获得的标签输出进行组合。

**Weighted\_Sum LSTM**: 我们首先使用每个单独的文本视角来训练LSTM分类器，然后使用加权和规则（Marler和Arora，2010）组合所有单视角LSTM分类器的获得的概率输出。

**Ensemble LSTM (Our approach)**: 我们提出的联合层方法。