利用深度学习融合模型提升文本内容安全的研究

汪少敏[1,2]，王铮[1,2]，任华[1,2]
（1. 移动互联网系统与应用安全国家工程实验室，上海 201315；
2. 中国电信股份有限公司研究院，上海 200122）

摘　要：互联网和移动互联网中的信息内容急速膨胀，导致其中充斥着违法违规和不良信息，影响互联网空间的内容安全。基于敏感词匹配的传统文本内容安全识别方法忽略上下文语义，导致误报率高、准确率低。在分析传统文本内容安全识别方法的基础上，提出了利用深度学习的融合识别模型以及模型融合算法流程。深入介绍了基于利用深度学习的融合识别模型的文本内容安全识别系统，并进行了实验验证。结果表明，所提模型可以有效解决传统识别方法缺乏语义理解造成误报率高的问题，提高了不良信息检测的准确性。

关键词：内容安全；违法违规和不良信息；深度学习；文本识别
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2020145

Research on fusion model based on deep learning for text content security enhancement

WANG Shaomin[1,2], WANG Zheng[1,2], REN Hua[1,2]
1. Mobile Internet System and Application Security National Engineering Laboratory, Shanghai 201315, China
2. Research Institute of China Telecom Co., Ltd., Shanghai 200122, China

Abstract: The rapid expansion of information content on the internet and the mobile internet has resulted in violations of laws and regulations and bad information, which affects the content security of the internet space. Traditional text content security recognition methods based on matching of sensitive words ignore context semantics, resulting in high false positive rate and low accuracy. Based on the analysis of traditional text content security recognition methods, a fusion recognition model using deep learning and a model fusion algorithm process were proposed. Text content security recognition system based on the fusion recognition model using deep learning and experimental verification was introducted deeply. Results show that the proposed model can effectively solve the problem of high false positive rate caused by the lack of semantic understanding of traditional recognition methods, and improve the accuracy of the bad information detection.

Key words: content security, illegal information and unhealthy information, deep learning, text recognition

1　引言

随着互联网和移动互联网的飞速发展，互联网上的信息内容急剧膨胀。各种社交网络、媒体平台产生的文本信息数量每天超过 5 亿条，视频、图片信息数量超过 10 亿条，并且这种趋势还在继

收稿日期：2020−03−12；修回日期：2020−04−22

续[1-2]。然而这些信息中充斥着大量违法违规和不良信息，包括极端言论、色情、暴力恐怖、欺诈等内容。这些违法违规和不良信息严重影响互联网空间的内容安全，从而造成国家安全受损和社会稳定隐患，对青少年产生不良影响[3]。因此，通过对违法违规和不良信息的识别过滤，提升互联网内容安全，达到营造清朗网络空间的目的，是目前业界关注的重要课题。

根据媒体类别的不同，互联网中的违法违规和不良信息主要分为文本、图片、视频 3 种类型。针对文本类的违法违规和不良信息，传统的识别方法为通过预先设置好的敏感词库，规则匹配从互联网采集的文本信息[4]，若该条信息中有匹配成功的敏感词或敏感词组合，则识别该条信息为违法违规和不良信息。由于同一单词或词组在不同语境和上下文中表达不同的观点，而传统的违法违规和不良信息识别方法只匹配单词或词组，而忽略了上下文语义，造成了识别的准确率低、误报率高，如某些文本信息包含敏感词，但整体的意思并不符合违法违规和不良信息；而某些文本信息并不包含敏感词或词组，但整体的意思是违法违规的，如果用传统识别方法识别这类信息，会造成误报和漏报。低准确率、高误报率不仅造成极大的人工审核成本，还造成互联网内容显示得不恰当，从而影响互联网空间的内容安全。

由于深度学习技术能准确地分析出非结构化信息的语义[5]，所以在传统的内容安全识别架构中引入深度学习技术，形成深度学习和敏感词匹配的融合模型。利用深度学习的融合模型对互联网中文本类信息进行识别过滤，弥补传统识别方法对上下文语义处理的缺失，能有效解决敏感词匹配造成的误报和漏报问题，提升识别违法违规和不良信息的准确率。

本文针对传统文本类内容安全识别方法的局限性，探讨了利用深度学习的融合模型识别文本内容安全的技术和系统应用，并进行了实验验证。通过实验数据验证了利用深度学习的融合模型能有效提高文本内容安全识别的准确率。此方法和系统将大幅降低文本内容安全识别过滤的人工审核成本，能更有效准确地过滤违法违规和不良信息，提升互联网空间的内容安全。

2 传统文本类内容安全识别方法

传统文本类内容安全识别系统一般由 3 个部分组成：敏感词库、信息采集、匹配算法，具体如图 1 所示。

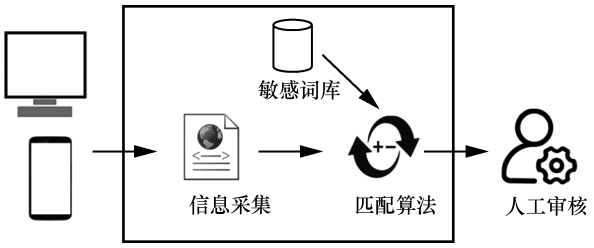

图 1 传统文本类内容安全识别系统

敏感词库一般依照相关法律、法规、行政要求、企业规章和标准规定制定，满足匹配过滤的需求。敏感词库分级分类构建，例如，某个敏感词库包含 4 级，每级包含若干分类，如反动、色情、暴力恐怖、邪教与迷信、欺诈、谩骂、违反道德等。敏感词库的内容以词语为主，辅以少量短语或句子。敏感词库还可以包含通过对词语进行与或非的关系组成的敏感词组合，以达到更准确过滤违法违规和不良信息的目的。

信息采集是将互联网和移动互联网中的信息内容统计到系统中，以进行下一步的信息分析。信息采集的范畴可以包括网页、App、小程序内容等。信息采集的手段有网页爬虫、App 拨测、服务器出口分光等[6]。信息采集到的内容载体有文本、图片、视频。其中可以直接对文本进行下一步的分析，图片中的文字可以通过 OCR（optical character recognition，光学字符识别）技术将其转换为文本[7]，再做分析。

匹配算法通过规则匹配敏感词库的方法，对信息采集后的文本类数据进行分析，判断其是否为违法违规和不良信息，以及属于哪一类违法违规和不良信息。使用较多的匹配算法是 DFA（deterministic finite automaton，确定有穷自动机）算法[8]，通过有限状态机的方式，在计算量不大的情况下，进行匹配。

传统文本类内容安全识别方法的关键在于对采集的文本信息进行词语或词组匹配，这种方式存在误报漏报的风险。例如，某网站页面出现语句“展示运动健儿们激情澎湃的时刻到了”，在分析从该网页采集的文本信息时，由于文本中“激情”一词和敏感词库中的色情类词汇“激情”匹配，系统判断该网页为色情类违法违规和不良信息，导致误判。若某些采集到的文本信息并不包含敏感词库中词语或词组，但整体的意思却是违法违规的，则会造成漏判。仅通过敏感词匹配来识别信息，缺失了对上下文语义的理解，造成传统文本类内容安全识别方法的准确率降低。

3　利用深度学习的融合识别模型

在文本类内容安全识别中引入基于深度学习的神经网络模型，解决传统文本类内容安全识别方法中的上下文语义理解缺失的问题。神经网络模型虽然能够对文本信息进行上下文关联的语义理解分类[9]，然而，由于神经网络模型的训练需要涵盖每个类别标签的大量样本数据，而违法违规和不良信息的类别标签存在时效性，会经常进行增删，在增加某一类别时，相应的敏感词库能够快速更新，而对应新类别标签的样本数据却不那么容易获得，所以无法做到样本库的及时有效更新。所以，仅依赖神经网络模型来识别违法违规和不良信息，存在漏报的风险。所以，应综合应用敏感词匹配和神经网络模型两种方法，结合神经网络和敏感词匹配两种方法的融合识别模型结构如图 2 所示。

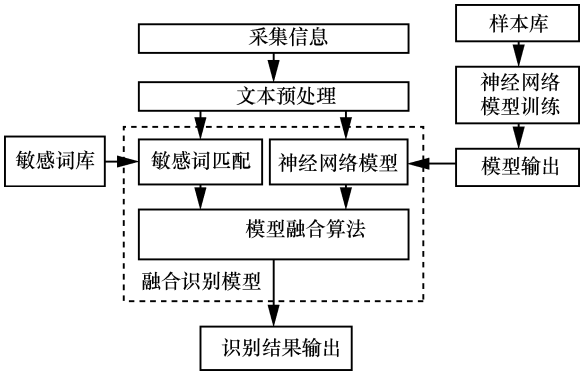

图 2　结合神经网络和敏感词匹配两种方法的融合识别模型结构

融合识别模型包含两种识别方法：敏感词匹配和神经网络模型，通过模型融合算法，发挥两种模型的不同优势，实现两种方法的有效融合。融合识别模型既实现了对上下文语义的分析，又能够在样本库内容不够完备的情况下，尽可能避免漏报，有效提高识别准确率。

对采集的文本信息进行分词等预处理后，敏感词匹配算法通过预先构建的敏感词库，进行敏感词匹配过滤；也可以通过对文本信息的分析转换，通过结构化的查询语言检索查询知识库或知识图谱，判断是否为违法违规或不良信息。基于深度学习的神经网络模型将文本信息进行分词、词向量化等预处理后，翻译成计算机能够识别的语言，输入已训练好的神经网络模型进行计算，得到神经网络模型预测的识别结果[9]。敏感词匹配和神经网络模型两种方法输出的识别结果通过模型融合算法来综合判断是否为违法违规或不良信息以及为哪一类违法违规或不良信息。

对于文本信息来说，模型融合算法流程如图 3 所示。

步骤 1　当神经网络模型输出为某个类别（标签类别中包含正常信息的类别）的概率值大于或等于 90%时，模型融合算法输出为神经网络模型的最大概率值类别。

步骤 2　当步骤 1 条件不满足时，若神经网络模型识别为某个类别的概率值大于或等于 80%，同时敏感词匹配识别结果为正常信息，模

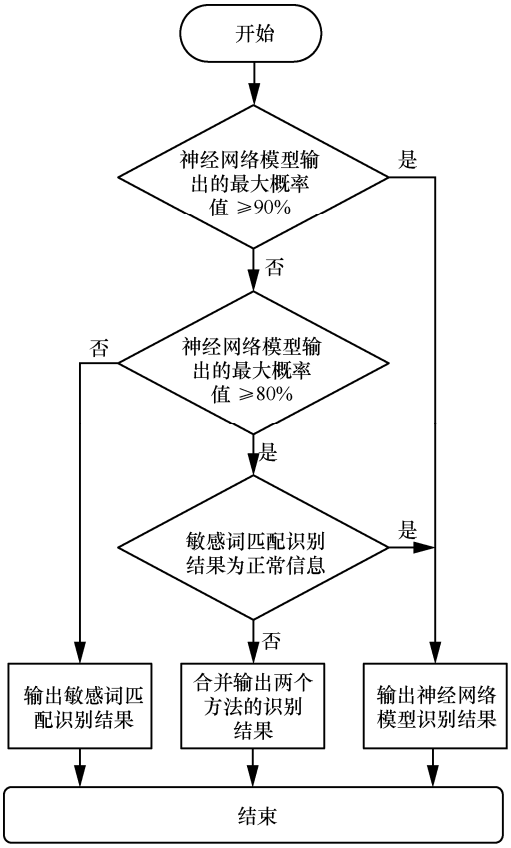

图 3　模型融合算法流程

型融合算法输出为神经网络模型的最大概率值类别。

步骤 3　当步骤 2 条件不满足时，若神经网络模型的识别结果的某个类别概率值大于或等于 80%，同时敏感词匹配识别结果为违法违规或不良信息，模型融合算法同时输出神经网络模型的最大概率值类别和敏感词匹配识别结果。

步骤 4　当步骤 3 条件不满足时，若神经网络模型的识别结果的最大概率值小于 80%，则模型融合算法输出为敏感词匹配识别结果。

4　文本内容安全识别系统

基于利用深度学习的融合识别模型的文本内容安全识别系统由数据采集模块、敏感词库模块、告警模块、统计分析模块和模型模块组成。

- 数据采集模块主要实现两个功能：一是通过网络爬虫、App 拨测、分光接口等手段，获取互联网和移动互联网中的文本信息和图片信息；二是若采集的是图片信息，通过 OCR 识别提取图片中的文本信息。
- 敏感词库模块主要实现词库的输入、更新和管理，包括敏感词的添加、删除、更新、备份和回滚。敏感词库应根据互联网环境和政策法规要求实时更新。
- 告警模块实现系统对发现的内容安全事件进行告警。
- 统计分析模块针对互联网中违法违规和不良信息的识别结果进行多维度的统计分析和图形化展示，如统计当日每小时的各类别的违法违规和不良信息数据；输入关键词，统计设定时间段内包含该关键词的违法违规和不良信息内容等。采用第 3 节所述的融合识别模型，对采集的文本信息进行检测，识别筛选出其中的违法违规和不良信息。
- 模型模块中的敏感词匹配算法采用 DFA 算法，DFA 算法在数据检索中能够大大地提升检索效率，所以这种算法常用于敏感词的过滤。模型模块中的神经网络模型对文本信息的处理分为 3 部分：文本预处理、词向量表示和深度学习模型。首先对采集到的文本信息进行预处理，包括分词、去除停用词和非中文信息，输出更易于机器处理的文本数据；然后对这些文本数据进行词向量表示，转换成神经网络模型能够训练和预测的形式；最后输入神经网络模型进行识别和分类。

考虑到对上下文关联的分析，神经网络采用 LSTM（long short-term memory，长短期记忆）网络。LSTM 网络是 RNN（recurrent neural network，循环神经网络）的特殊类型，它利用 RNN 中上下层的动态相关性，学习长期依赖信息，所以 LSTM 对文本中的上下文关系有较好的

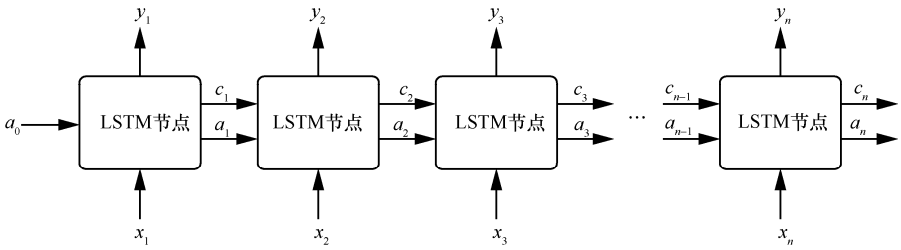

图 4　LSTM 网络结构

理解[10]。LSTM 网络结构如图 4 所示，网络结构中的 LSTM 节点有两个传输状态：a_n 和 c_n，相对于 RNN 结构，增加了 c_n 来表示节点的当前状态。节点中增加的门控结构实现了长序列运算中的选择记忆能力，使得 LSTM 继承了 RNN 的上下文理解能力，在长记忆方面也有更好的表现，更适用于文本数据的分类分析[11]。

5　实验验证

基于本文中的融合识别模型和应用该模型的文本内容安全识别系统，搭建了实验系统进行验证。实验系统的敏感词匹配算法采用 DFA 算法，神经网络模型采用 TensorFlow 框架实现。神经网络模型采用 LSTM 网络，网络结构为隐藏层 1 层、全连接层 1 层。每次训练的 batch（批次）大小为 64，词向量维度为 150 维，序列长度指定为 140 个词语。输出的类别为 4 类：国家安全、色情、欺诈与广告、正常。神经网络的训练样本库有 6 万条数据，其中正向样本为 3 万条，反向样本为 3 万条。互联网数据的信息采集是通过对天翼阅读 App 进行拨测截屏抓取，再对截屏图片进行 OCR 识别，将其转换为文本信息，输入融合识别模型中进行识别分类。

实验按句子颗粒度、截屏颗粒度，分别比较传统敏感词匹配方法和利用深度学习融合模型的分类准确率。句子颗粒度是指对每个句子进行识别，判断是否为违法违规或不良信息；截屏颗粒度是指对每个截屏进行识别，判断是否包含违法违规或不良信息。验证的方法为系统给出识别结果后，人工逐条判断识别结果是否正确。实验分为 5 次进行，以句子颗粒度验证时，每次采集的数据为 200 个句子，同时进行敏感词匹配和融合识别，得到两种方法对句子颗粒度的识别准确率；以截屏颗粒度验证时，每次采集 100 个截屏，同时进行敏感词匹配和融合识别，得到两种方法对截屏颗粒度的识别准确率。验证结果如图 5 所示。

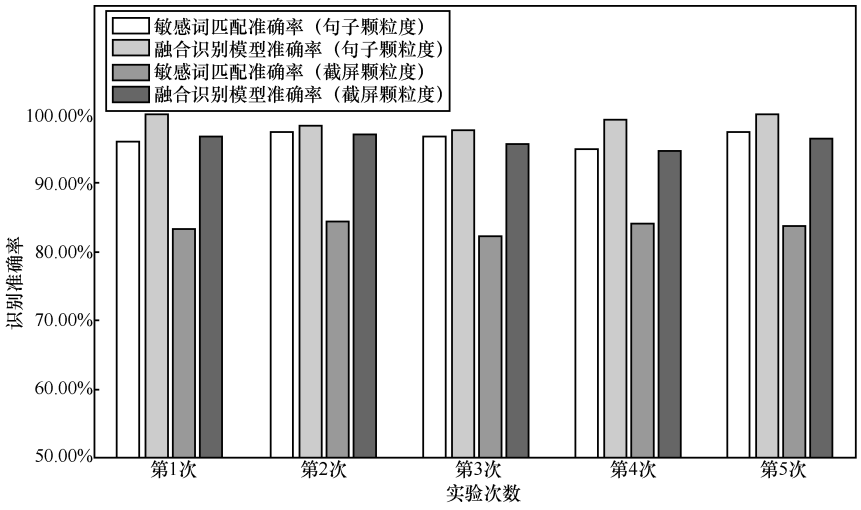

图 5　实验验证数据

从图 5 中可以看出，利用深度学习的融合识别模型对违法违规和不良信息的识别准确率明显高于传统的敏感词匹配方法。

6 结束语

本文在分析传统文本内容安全识别方法的基础上，针对传统文本内容安全识别方法缺失上下文语义理解的缺陷，提出了利用深度学习技术的融合识别模型，还阐述了应用融合识别模型构造的文本内容安全识别系统，并进行了实验验证。本文提出的利用深度学习技术的融合识别模型，有效解决了传统文本内容安全识别方法中语义理解缺失的问题，提高了不良信息检测的准确性，在减少人工成本的同时，实现准确有效的智能检测，提升互联网空间的内容安全，为打造清朗网络空间助力。

参考文献：

[1] ZHAO Z Y, LI M, LI C, et al. Dietary preferences and diabetic risk in China: a large-scale nationwide internet data based study[Z]. 2019.

[2] 高显俊, 黄儒乐. 互联网数据在高校大数据平台中的应用研究[J]. 科技资讯, 2019, 17(36): 12-13, 15.
GAO X J, HUANG R L. Research on the application of internet data in big data platform of universities[J]. Science & Technology Information, 2019, 17(36): 12-13, 15.

[3] 刘丹. 基于信息贫困理论的青少年信息行为浅析[J]. 时代金融, 2020(3): 138-140.
LIU D. An analysis of youth information behavior based on information poverty theory[J]. Times Finance, 2020(3): 138-140.

[4] BILLE P. Fast searching in packed strings[J]. Journal of Discrete Algorithms, 2010, 9(1).

[5] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[6] 邹一心, 范海平. 爬虫技术在 WAP 网站内容监测中的应用[J]. 电信科学, 2010, 26(Z1): 164-166.
ZOU Y X, FAN H P. Application of reptile technology in content monitoring of WAP website[J]. Telecommunications Science, 2010, 26(Z1): 164-166.

[7] GATOS B. A binary-tree-based OCR technique for machine-printed characters[J]. Engineering Applications of Artificial Intelligence, 1997, 10(4).

[8] BALCÁZAR J L, DÍAZ J, GAVALDÀ R, et al. The query complexity of learning DFA[J]. New Generation Computing, 1994, 12(4): 337-358.

[9] 山世光, 阚美娜, 刘昕, 等. 深度学习: 多层神经网络的复兴与变革[J]. 科技导报, 2016, 34(14): 60-70.
SHAN S G, KAN M N, LIU X, et al. Deep learning: the revival and transformation of multi layer neural networks[J]. Science & Technology Review, 2016, 34(14): 60-70.

[10] 汪少敏, 杨迪, 任华. 基于深度学习的文本分类系统关键技术研究与模型验证[J]. 电信科学, 2018, 34(12): 117-124.
WANG S M, YANG D, REN H. Key technology research and model validation of text classification system based on deep learning[J]. Telecommunications Science, 2018, 34(12): 117-124.

[11] 蔡鑫, 娄京生. 基于 LSTM 深度学习模型的中国电信官方微博用户情绪分析[J]. 电信科学, 2017, 33(12): 136-141.
CAI X, LOU J S. Sentiment analysis of telecom official micro-blog users based on LSTM deep learning model[J]. Telecommunications Science, 2017, 33(12): 136-141.

[作者简介]

汪少敏（1983– ），女，移动互联网系统与应用安全国家工程实验室、中国电信股份有限公司研究院高级工程师，主要研究方向为内容安全识别技术、人工智能技术和自然语言处理。

王铮（1973– ），男，移动互联网系统与应用安全国家工程实验室、中国电信股份有限公司研究院高级工程师，主要研究方向为信息安全、人工智能技术、大数据架构和数据挖掘分析。

任华（1977– ），女，移动互联网系统与应用安全国家工程实验室、中国电信股份有限公司研究院高级工程师，主要研究方向为内容信息安全、数据分析和人工智能技术。