重大突发疫情事件中的谣言识别

黄哲英 南开大学； 刘勘 中南财经政法大学

论文摘要

新冠疫情爆发以来，相关谣言肆虐传播，但传统的谣言识别模型却难以有效判别疫情谣言，因为相较于大量历史谣言数据，疫情谣言的数量还不足以训练出良好的分类器。因此，建立一个以少量谣言数据为基础的疫情谣言识别模型紧迫且重要。针对训练数据量不足的问题，利用文本增强和生成对抗网络方法，生成大量与疫情谣言相似的谣言数据，达到提高疫情谣言鉴别效果的目的。首先分析疫情谣言的文本特征，提取能表征疫情谣言的特征词；然后基于生成对抗（GAN）思想，构建疫情谣言生成模型，将不含疫情谣言特征的历史谣言，利用疫情谣言特征词库进行文本增强，并生成大量含有疫情谣言特征的新谣言数据；最后，在疫情谣言中补充新生成的谣言数据，从而训练出更准确的疫情谣言分类模型。实验表明，使用GAN扩充训练集后，识别效果提高三个百分点，明显优于传统机器学习和深度学习算法，为重大突发疫情事件中谣言的识别提供了新的途径。

系统模型

1. 疫情谣言主题特征

表1 疫情谣言主题词汇表

主题	词1	词2	词3	词4	词5
疾病防治	设施	抢救	医疗	病毒	聚集
灾害救助	药物	捐赠	支援	调用	无偿
政策解读	违法	公布	复工	调任	解封
人物聚焦	英雄	演讲	医生	行程	重症
民生保障	交通	小区	快递	拦截	消毒

2. 疫情谣言文本特征

图1 疫情谣言情感词云图

表2 疫情谣言词语频数表

词语	次数	词语	次数
不要	48	医院	60
重	44	消毒	34
……	……	……	……
极	28	酒精	27
千万	26	药	19
一定	16	防治	17
非常	11	专家	10
……	……	……	……

表3 疫情谣言关键词表

词语	Sum(TF-IDF)	词语	Sum(TF-IDF)
美国	18.22	肺炎	8.80
中国	13.85	医院	8.66
医疗	13.43	日本	8.21
疫情	10.45	俄罗斯	8.16
允恶	10.16	封城	7.90
开学	9.71	消毒	7.82
武汉	9.44	……	……

3. 谣言文本增强

表4 疫情谣言词库表

程度词表		情绪词表		领域词表		主题词表	
原词	替换	原词	替换	原词	替换	原词	替换
很	极	坏	阴险	疾病病情	疫情	设备装置	设施
较为	极为	粗鲁	野蛮	治疗预防	防治	急救急诊	抢救
愈加	更为	生气	暴怒	非典SARS	新冠	药品药	药物
要	一定要	危险	惊险	帮助救助	支援	高铁公交	交通
少	极少	可怕	恐怖	筹款借款	捐赠	勇士伟人	英雄
多	极多	放纵	嚣张	上班开学	复工	犯法作案	违法
……	……	……	……	……	……	……	……

前半部分通过对疫情谣言主题与文本特征的分析，得到主题词、情感词、高频词及关键词四方面的特征，据此构建疫情谣言词库，对历史谣言文本进行词汇转换，实现文本增强效果。后半部分基于GAN构建疫情谣言生成模型，生成大量含有疫情谣言特征的新谣言数据，利用原有数据与新谣言数据，得到最终的疫情谣言分类模型，实现谣言鉴别功能。

4. 基于GAN的疫情谣言识别模型

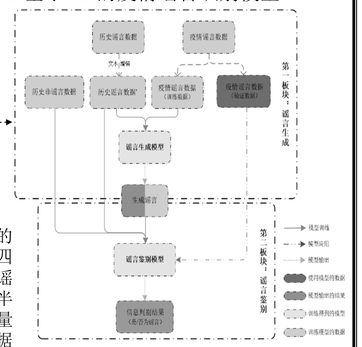

图2 基于GAN的疫情谣言识别流程图

论文简介

2020年1月中下旬开始，新型冠状病毒肺炎疫情逐渐严峻，一些造谣者在网上大肆发布疫情相关的谣言信息。本文以提供可靠的疫情谣言鉴别模型为最终目的，实现对重大突发疫情事件中相关信息的真伪鉴别。其中的核心问题在于，与历史谣言数据相比，疫情谣言的数据量相对较少，且疫情谣言有其独特特征，难以利用现有的判别模型进行鉴别。对此，本文从文本增强和生成对抗两个方面构建疫情谣言判别模型。首先基于疫情谣言的文本特征构建疫情谣言词库，从而对历史谣言进行符合疫情谣言特征的文本增强，再建立基于GAN的疫情谣言生成模型，以此来扩充疫情谣言的数据量，实现更加精准的疫情谣言鉴别效果。

算法原理

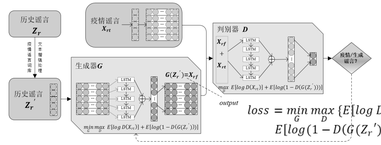

图3 疫情谣言生成模型图

生成器将历史谣言Z_r'转变成带有疫情谣言特征的生成谣言$G(Z_r')$，判别器区分真实的疫情谣言和生成器生成的谣言。当模型迭代足够的次数后，判别器的准确率趋近于50%，疫情谣言和生成谣言具有较高的相似性，判别器无法辨别数据来源。此时生成谣言$G(Z_r')$便可以用来扩充训练数据集，作为标注数据展开模型训练。

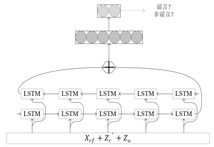

图4 疫情谣言鉴别模型结构图

生成大量疫情谣言之后就可以构建分类器实现疫情谣言鉴别。本文使用基于BiLSTM网络的分类器模型。

实验仿真

实验1：与传统的机器学习和深度学习方法对比实验。分别使用传统机器学习算法NB、SVM、DT，集成学习算法XGBoost以及BP算法进行训练，与深度学习中的CNN、RNN、LSTM和BiLSTM进行比较。

表5 实验1的结果对比表

序号	模型	Acc (train)	Loss	Acc (test)	NCoV
1	NB	——	——	0.9557	0.3082
2	SVM	——	——	0.9677	0.5205
3	DT	——	——	0.7644	0.2562
4	XGBoost	——	——	0.9136	0.4534
5	BP	——	——	0.9565	0.5014
6	CNN	0.9878	0.0347	0.9658	0.7466
7	RNN	0.9676	0.0848	0.9668	0.7260
8	LSTM	0.9801	0.0532	0.9793	0.8082
9	BiLSTM	0.9830	0.0432	0.9807	**0.8151**
10	GAN+BiLSTM (LSTM+LSTM)	0.9743	0.0704	0.9520	**0.8288**

表6 实验2的结果对比表

序号	模型	Acc-test (before)	NCoV (before)	Acc-test (after)	NCoV (after)
1	NB	0.9557	0.3082	0.9622	0.3096
2	SVM	0.9677	0.5205	0.9705	0.4671
3	DT	0.7644	0.2562	0.8048	0.2753
4	XGBoost	0.9136	0.4534	0.9170	0.3699
5	BP	0.9565	0.5014	0.9625	0.5068
6	CNN	0.9658	0.7466	0.9687	0.7808
7	RNN	0.9668	0.7260	0.9745	0.8082
8	LSTM	0.9793	0.8082	0.9809	0.8151
9	BiLSTM	0.9807	**0.8151**	0.9816	**0.8356**
10	GAN+BiLSTM (LSTM+LSTM)	0.9520	**0.8288**	0.9775	**0.8493**

实验2：文本增强前后效果对比实验。对文本进行增强处理后，使用实验1中的算法再次进行实验，检验文本增强处理在此研究中的应用效果。

实验3：GAN的生成器与判别器使用不同算法进行对比实验。分别使用LSTM与BiLSTM这两种算法作为生成器与判别器的核心算法进行对比，选择最优生成模型。

表7 实验3的结果对比表

序号	模型	Acc (train)	Loss	Acc (test)	NCoV
1	GAN+BiLSTM (BiLSTM+LSTM)	0.9783	0.0588	0.9766	0.8356
2	GAN+BiLSTM (LSTM+LSTM)	0.9790	0.0577	0.9775	**0.8493**
3	GAN+BiLSTM (BiLSTM+BiLSTM)	0.9860	0.0380	0.9803	0.8219
4	GAN+BiLSTM (LSTM+BiLSTM)	0.9768	0.0568	0.9768	0.8014

表8 实验4的结果对比表

序号	模型	Acc (train)	Loss	Acc (test)	NCoV
1	GAN+CNN (LSTM+LSTM)	0.9738	0.0729	0.9641	0.8014
2	GAN+RNN (LSTM+LSTM)	0.9603	0.1051	0.9638	0.7808
3	GAN+LSTM (LSTM+LSTM)	0.9762	0.0630	0.9771	0.7397
4	GAN+BiLSTM (LSTM+LSTM)	0.9790	0.0577	0.9775	**0.8493**
5	GAN+CNN (BiLSTM+LSTM)	0.9732	0.0726	0.9691	0.6301
6	GAN+RNN (BiLSTM+LSTM)	0.9518	0.1293	0.9512	0.7260
7	GAN+LSTM (BiLSTM+LSTM)	0.9762	0.0627	0.9763	0.7945
8	GAN+BiLSTM (BiLSTM+LSTM)	0.9783	0.0588	0.9766	**0.8219**

实验4：GAN训练完成后，连接不同分类器的对比分析实验。分别用GAN+CNN以及RNN、LSTM、BiLSTM进行结果对比，选择最优的鉴别模型。

根据实验结果可以发现，经过文本增强处理与生成对抗扩充数据后，模型在疫情谣言上得到的预测效果，比之前最优值提高三个百分点。当使用LSTM构造生成器与判别器，并使用BiLSTM为鉴别模型的算法时，模型效果最优。可见，本文提出的模型可以在疫情谣言的鉴别中定向提高模型准确率。

论文结论

疫情当前，为了使谣言对社会的危害降到最低，就要争取及早准确地鉴别疫情相关谣言。为了弥补疫情数据的不足，实现更加精准的疫情谣言鉴别，本文一方面构建疫情谣言词库，利用其进行文本增强；另一方面使用GAN对此次疫情的相关谣言进行特征提取，将历史谣言转化为具有疫情谣言特征的生成谣言，获得大量与此次疫情相关联的谣言数据。使用增强后的训练数据集进行判别模型训练，提高了谣言判别的准确率。

随着时间推移，获得证实的疫情谣言数量越来越多，此时，可以对疫情谣言数据集进行补充更新，使得谣言生成模型更加准确地学习疫情谣言的特征，不断提高生成数据与真实疫情谣言之间的相似度，获得更高质量的生成谣言，进而达到更加精准的谣言鉴别效果。

本文的研究方法是针对疫情数据提出的，不论是此次疫情还是未来可能会出现的疫情，本文构建的模型都可以对重大突发疫情的谣言治理起到辅助作用，为普通网民提供相应的判别依据。但疫情毕竟是极少发生的，对疫情谣言及传播者的特征提炼仍有难度，如果加入更多的特征，如谣言传播主体的行为特征，有可能继续提高谣言判别的效果。此外，现实中谣言与非谣言的数据不平衡现象明显，如何在研究中加入对现实情况的考量，也需要展开进一步的研究。