فصلنامه سیستم‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)

Summer 2021, 2 (2), 9-16
DOR: 20.1001.1.27832570.1400.2.3.2.7

Received: 21 Mar. 2021
Accepted: 25 May 2021

Adversarial Attacks on a Text Sentiment Analysis Model

Sahar Mokarrami Sefidab[1], Seyed Abolghasem Mirroshandel[2*], Hamidreza Ahmadifar[3], Mehdi Mokarrami Sefidab[4]

1. Faculty of Engineering, University of Guilan, Rasht, Iran. sahar.mokarrami94@gmail.com
2.Assistant Professor, Faculty of Engineering, University of Guilan, Rasht, Iran. *(Corresponding Author)* mirroshandel@guilan.ac.ir
3.Assistant Professor, Faculty of Engineering, University of Guilan, Rasht, Iran. ahmadifar@guilan.ac.ir
4. Payam-e Noor University of Guilan, Rasht, Iran. info.mokarrami@gmail.com

Abstract

Background and Purpose: Recently some researchers have shown that deep learning models, despite their high accuracy, can be vulnerable through some manipulations of their input samples. This manipulation leads to the production of new samples called Adversarial examples. These samples are very similar to the original ones, so humans cannot differentiate between these samples and the original, and cannot remove them from the dataset before predicting the model and preventing model errors. Various types of research have been done to generate malicious samples and inject them into the model, among which, the production of text samples has its own difficulties due to the discrete nature of the text. In this research, we tried to reach the highest level of vulnerability by providing a method with the least manipulation of the input data, and by testing the proposed method, we were able to bring the accuracy of CNN and LSTM models to less than 10%.

Methods: In this research, for making malicious samples, first, a word that can increase the amount of error in the classification prediction is selected from the word dictionary as a candidate word for replacement by using Taylor expansion and then considering the importance of each word in the calculated cost of the corresponding candidate word, we proposed an arrangement for substitution between words. Finally, we moved the words in the specified order until the output of the model changed.

Results: The evaluation of the presented method on two sentiment analysis models, LSTM and CNN, has shown that the proposed method has been very effective in reducing the accuracy of both models to less than 10% with a small number of replacements and this indicates the success of the proposed method compared to some other similar methods.

Conclusion: As mentioned, most of the attention of science and industry is on the production of different systems using deep learning methods, so their security of them is also important. It is important to increase the strength of the models against adversarial examples. In this research, a method with the least amount of manipulation was presented to produce textual conflict samples. It seems that in the future it will be possible to use different methods of making natural texts to produce samples that, in addition to the apparent similarity to the original sample, are also comprehensible in terms of content.

Keywords: Text Attacks, Adversarial Examples, Loss function gradient, Sentiment Analysis, Natural Language Processing.

فصلنامه سیستم‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)
انجمن علمی تجارت الکترونیکی ایران

تاریخ دریافت: ۱۳۹۹/۱۲/۰۲
تاریخ پذیرش: ۱۴۰۰/۰۲/۲۵

سال دوم، تابستان ۱۴۰۰
شمارهٔ دوم، صص: ۹ – ۱۶

حملات تخاصمی در یک مدل تحلیل احساس متن

سحر مکرمی‌سفیدآب[۱]، سیّدابوالقاسم میرروشندل[۲*]، حمیدرضا احمدی‌فر[۳]، مهدی مکرمی‌سفیداب[۴]

۱. دانشکده فنی، دانشگاه گیلان، رشت، ایران. sahar.mokarrami94@gmail.com
۲. استادیار، دانشکده فنی، دانشگاه گیلان، رشت، ایران.(نویسنده مسئول) mirroshandel@guilan.ac.ir
۳. استادیار، دانشکده فنی، دانشگاه گیلان، رشت، ایران،ahmadifar@guilan.ac.ir.
۴. دانشگاه پیام‌نور واحد رشت، رشت، ایران. info.mokarrami@gmail.com

چکیده: شبکه‌های عصبی عمیق دقت و کارایی بالایی در حل مسائل مختلف دارند، اما در برابر نمونه‌های تخاصمی آسیب‌پذیر هستند. این د سته از نمونه‌های مخرب به‌منظور فریب مدل آموزش‌دیده و برر سی آ سیب‌پذیری مدل‌های شبکه ع صبی تولیدمی شوند. در حوزه متن روش‌های موفق زیادی برای ساخت این نمونه‌ها ارائه‌نشده‌است. در این پژوهش یک روش قوی مبتنی‌بر گرادیان تابع هزینه مدل برای تولید نمونه‌های تخاصمی متنی ارائه‌شده و نشان‌داده‌است که می‌توان با جایگزینی تعداد کمی از کلمات موجود در نمونه‌های اصلی با کلماتی که بیشترین تأثیر منفی را روی تصمیم طبقه‌بند دارند، نمونه‌های جدیدی مشابه با نمونه‌های اولیه برای فریب طبقه‌بند تحلیل احساس در سطح کلمه تولیدنمود. در نهایت با بهره‌گیری از این نمونه‌ها دقت دو مدل طبقه‌بند از پیش‌آموزش‌دیده، برر سی شد. روش مورد استفاده در این پژوهش، با دست‌کاری اندک نمونه‌های ورودی، موفق به کاهش دقت طبقه‌بندی از ۸۶ درصد به کمتر از ۱۰ درصد شده‌است.

واژه‌های کلیدی: حملات متنی، نمونه‌های تخاصمی، گرادیان تابع هزینه، تحلیل احساس، پردازش زبان طبیعی.

۱. مقدمه

در دنیای امروز رویکردهای یادگیری عمیق تقریبا در تمامی حوزه‌های هوش مصنوعی به‌خصوص تجزیه و تحلیل احساس، توانسته‌اند با ارائه مدل‌های بهینه، راه‌حل‌های سریع و دقیقی را برای حل مسائل مختلف داشته‌باشند. درحالی‌که برخی مطالعات نشان‌داده‌است که مدل‌های یادگیری عمیق در برابر نمونه‌های تخاصمی به‌شدت آسیب‌پذیر هستند به‌طوری که با تشخیص اشتباه خسارات جبران نا پذیری را تحمیل‌می‌نمایند. نمونه‌های تخاصمی، نوع تغییریافته و دست‌کاری‌شده نمونه‌های ورودی اولیه خود هستند که باعث فریب مدل‌های شبکه عصبی و رخداد خطا در آن‌ها می شوند [۱]. این قبیل نمونه‌ها شباهت بسیار زیادی با نمونه‌های اولیه خود دارند به همین دلیل انسان نمی‌تواند بین این نمونه‌ها با نمونه‌های اصلی موجود در مجموعه دادگان تفاوتی قائل شود؛ بنابراین نمی‌تواند قبل از پیش‌بینی مدل آن‌ها را از مجموعه داده‌ها حذف‌نماید تا مانع از خطای مدل در پیش‌بینی شود. به‌همین‌منظور باید به‌دنبال روشی برای مبارزه با این نمونه‌ها و مقاوم‌سازی مدل‌های شبکه عصبی دربرابر آن‌ها بود.

تلاش‌های مختلف برای ساخت نمونه‌های تخاصمی از یک سو و استفاده از این نمونه‌ها در مقاوم‌سازی و بررسی آسیب‌پذیری مدل‌های یادگیری عمیق از سویی دیگر منجر به گشایش دریچه جدیدی از دانش در حوزه یادگیری عمیق شده‌است. با استفاده از این نمونه‌ها می‌توان به بررسی امنیت مدل‌های شبکه عصبی پرداخته و آسیب‌پذیری مدل‌های مختلف متنی را نشان‌داد. از طرفی با یافتن محدودیت‌های مدل می‌توان ارزیابی و تفسیر مناسبی از آن مدل داشت. مثلاً در پژوهشی از این نمونه‌ها برای ارزیابی مدل‌های درک مطلب متنی به‌منظور تو سعه مدل‌های دقیق‌تر [۲] و در پژوه شی دیگر برای مشاهده و سنجش میزان آسیب‌پذیری مدل‌های ترجمه ماشینی استفاده‌شده‌است [۳].

در زمینه تصویر پژوهش‌های بسیاری برای ساخت نمونه‌های تخاصمی ارائه‌شده‌اند که اکثراً موفق بوده‌اند. اما در حوزه متن این پژوهش‌ها موفقیت چندانی نداشته‌اند چراکه متن برخلاف تصویر ماهیتی گسسته دارد. هر تصویر از تعداد زیادی پیکسل تشکیل‌شده‌است که اعمال تغییرات کوچک روی آن‌ها و ایجاد اختلال در آن‌ها از دید انسان پنهان مانده و درعین‌حال می‌تواند مدل را نیز در تشخیص، گمراه‌نماید. اما در حوزه متن جابه‌جایی و تغییر در سطح کلمه یا حرف منجر به تغییر اساسی در ساختار متن می‌شود و مفهوم متن را به‌هم‌می‌ریزد. بنابراین استفاده مستقیم از روش‌های تولید نمونه‌های تخاصمی تصویری برای ایجاد اختلال و دست‌کاری داده‌های متنی کارآمد نیست [۴].

بر همین اساس روش‌های مختلفی برای ساخت نمونه‌های تخاصمی و ایجاد حمله روی مدل‌های متنی ارائه‌شده‌اند. این روش‌ها به دو نوع کلی حملات جعبه سیاه و جعبه سفید دسته‌بندی‌می‌شوند. در حملات جعبه سیاه، مدل شبکه عصبی و گرادیان آن به‌صورت کامل در دسترس نیست. بنابراین در این نوع از حملات از روش‌های آزمون و خطا برای تولید نمونه مخرب ا ستفاده‌می شود به‌گونه‌ای که پس از تولید نمونه‌ای خروجی مدل را بررسی‌نموده تا زمانی‌که نمونه مخرب مناسب یافت شود، اما برخلاف حملات جعبه سیاه در حملات جعبه سفید مهاجم به مدل شبکه عصبی دسترسی کامل دارد بنابراین در این قبیل حملات از روش‌های مبتنی‌بر گرادیان به‌عنوان یک رویکرد حمله استفاده‌شده و پژوهشگران با روش‌های مختلفی تلاش‌می‌کنند تا داده‌ها را به‌گونه‌ای‌که گرادیان تابع هزینه افزایش‌یابد، دست‌کاری‌نمایند و در نتیجه منجر به خطای مدل در پیش‌بینی شوند [۵]. حملات تخاصمی از نظرگاهی دیگر به دو صورت حملات هدفمند و حملات بی‌هدف نیز تقسیم‌بندی می‌شوند. در حملات هدفمند در جستجوی روشی هستیم تا مدل تقریباً به‌ازای بیشتر نمونه‌ها یک برچسب هدف را پیش‌بینی‌نماید درحالی‌که در حملات بدون‌هدف تنها از مدل می‌خواهیم تا در تشخیص برچسب نمونه‌ها دچار خطا شود [۶].

برقراری غلط های املایی روی کلمات، حذف، اضافه‌کردن و یا جایگزینی کلمات با کلمات مشابه دیگر، از جمله اولین کار های انجام‌شده برای تولید نمونه‌های تخاصمی در حوزه متن هستند. به‌عنوان نمونه، در یک حمله در سطح کاراکتر، یک متد ساده به‌صورت انتخاب یک کاراکتر اطراف کاراکتر مورد نظر در ساختاری مانند ساختار صفحه کلید و جایگزینی آن با کاراکتر اصلی معرفی‌شد. این مدل روشی برای حمله روی یک سیستم ترجمه ما شینی محسوب‌می شود. به‌طور مشابه در رو شی دیگر، نمونه‌هایی برای فریب Google Perspective API از طریق ایجاد برخی غلط املایی روی لغات ورودی یا اضافه‌کردن نقطه یا کاراکترهای مشابه به حروف تولیدشد. در این روش‌ها اغلب اوقات نمونه‌های تولیدشده از دید انسان پنهان نیست [۳،۷]. راهکار دیگری برای جایگزینی کلمات، ا ستفاده از الگوریتم ژنتیک ا ست که به صورت تکراری جمعیتی از راه حل های کا ند ید را به راه حل بهتر تکامل‌می‌دهد. بدین شکل که کلمه‌ای انتخاب شده و سپس با ا ستفاده از فاصله اقلیدسی، تعدادی از نزدیک‌ترین همسایه‌های آن کلمه انتخاب‌می شود و سپس کلمه‌ای که احتمال برچسب هدف برای فریب مدل را افزایش‌دهد به‌عنوان کلمه مورد نظر برای جایگزینی انتخاب می شود [۸]. یک روش دیگر، وابسته به گرادیان تابع هزینه برای تولید نمونه‌های تخاصمی است. در این روش ابتدا گرادیان تابع هزینه مدل به ازای تمامی کلمات موجود در متن، محاسبه‌می‌شود و پس از آن کلمه‌ای با بی شترین مقدار گرادیان به‌عنوان کلمه منا سب برای افزودن اختلال به آن و گمراهی مدل انتخاب‌می‌شود زیرا می‌تواند خطای مدل را افزایش‌دهد [۹]. در روش دیگری یک ا ستخری از کلمات مترادف، از نظر ظاهری مشا به و نیز غلط های املایی به‌ازای کلمات موجود، درنظرگرفته شده‌ا ست و جایگزینی کلمات موجود در نمونه با هرکدام از کلمات مشابه استخر بر اساس اهمیت هر کلمه با محاسبه گرادیان نسبت به ورودی انجام‌گرفته‌است [۱۰]. در پژوهشی دیگر فضایی برای تعبیه کلمات درنظرگرفته شده و بردار اختلال کوچکی در ف ضای تعبیه

برداری به کلمات موجود در آن فضا افزوده شده‌است و سپس این بردار جدید به نزدیک‌ترین کلمه اطراف خود در آن فضای برداری نگاشت‌شده‌است [۱۱]. در یکی از جدیدترین پژوهش‌ها برای تولید نمونه‌هایی با ساختار متنی مناسب، از یک مدل زبانی پوششی استفاده‌شده‌است. در این روش با نمونه‌گیری از توزیع احتمالاتی کلمات موجود در نمونه های اصلی که با کمک مدل زبانی پوششی محاسبه‌شده‌اند و به‌روزرسانی وزن‌های مدل زبانی مبتنی‌بر خروجی طبقه‌بند، طی چند گام نمونه جدیدی مشابه با نمونه اولیه تولیدشد [۴]. از سوی دیگر می‌توان برای تولید نمونه مناسب، جهت بردارهای لغاتی را که در فضای برداری قرارمی‌گیرند، محدود به کلمات موجود در آن فضا نمود [۱۲].

نوع دیگری از حملات به‌عنوان حمله سرا سری محسوب‌می شوند. این نوع از حملات مستقل از داده‌های ورودی هستند به‌گونه‌ای‌که بدون دست‌کاری و تغییر مستقیم نمونه‌های ورودی می‌توان یک اختلالی را به تمامی نمونه‌ها افزود تا بهازای بیشتر نمونه‌ها مدل دچار اشتباه در تشخیص شود [۱۳]. برای ایجاد حملات سرا سری متنی، در یکی از اولین روش‌ها، ابتدا تعدادی کلمه با طول مشخص به‌عنوان اختلال اولیه به ابتدای تمامی نمونه‌های اصلی اضافه‌شده، سپس برای این‌که اختلال خاصیت تخریب‌پذیری داشته‌باشد، گرادیان تابع هزینه مدل نسبت به هرکدام از کلمات اختلال به بردار کلمات افزوده‌می‌شود. در انتها نیز نزدیک‌ترین بردار به بردار قبلی در فضای تعبیه برداری به‌عنوان کلمه جایگزین انتخاب‌می‌شود [۱۳].

در یک پژوهش، برای اینکه نمونه‌های تخاصمی سراسری تولیدشده ساختاری مشابه با عبارت و جمله داشته و بامعنا باشند، از یک خودرمزگذار استفاده‌شده‌است. در این روش یک بردار نویز گاوسی انتخاب شده و به‌عنوان ورودی، به یک مولد داده‌می شود و مولد آن‌را به یک فضای پنهان منتقل‌می‌کند در نهایت با استفاده از رمزگشا خروجی فضای پنهان به یک عبارت بامعنی تبدیل‌می‌شود. در هربار تکرار این عملیات، تصویر گرادیان تابع خطای مدل حول بردار نویز، محاسبه‌شده و به نویز قبلی اضافه‌می شود که این کار برای به‌روزرسانی بردار نویز و افزایش خطای مدل انجام‌می‌گیرد [۱۴].

در روشی، برای تولید نمونه‌هایی که از نظر محتوایی مشابه با نمونه‌های اولیه هستند، ارزش هر کلمه در متن و نیز اختلاف مقادیر احتمالات برچسب نمونه اصلی با نمونه تغییریافته، محاسبه‌شده و به‌صورت حریصانه، کلمات داخل متن با کلمات مترادف خود جایگزین‌می شوند [۱۵]. در روشی دیگر بهترین جایگزینی با محاسبه گرادیان تابع هزینه نسبت به آن جایگزینی و انتخاب بیشترین مقدار گرادیان، مشخص شده و درنهایت با استفاده از جستجوی پرتویی یک دسته از بهترین کلمات برای فریب مدل و جایگذاری در متن ورودی انتخاب‌می‌گردد [۱۶،۱۷]. در یک روش دیگر از الگوریتم متروپلیس هستینگز برای تولید کلمات مناسب استفاده شده‌است. البته در برخی حالات استفاده از این روش منجر به تولید جملات و عبارات ناقص شده‌است [۵،۱۸].

در این پژوهش روشی قوی و وابسته به گرادیان تابع هزینه به‌صورت هدفمند به‌منظور ایجاد تغییراتی روی داده های ورودی ارائه‌می‌نماییم و نشان‌خواهیمداد که با استفاده از این روش تنها با جایگزینی تعداد بسیار کمی از کلمات موجود در نمونه‌های ورودی با کلماتی دیگر، می‌توان دقت طبقه‌بند را به‌شدت کاهش‌داد.

روش ارائه شده به صورت جایگذاری کلمات است که در آن هزینه هر جابه‌جایی به‌همراه ارزش هر کلمه در نمونه ورودی برای ایجاد یک حمله متنی مناسب درنظرگرفته‌می‌شود. درواقع برای ترتیب‌دادن یک حمله قدرتمند، با استفاده از بیشترین هزینه هر جابه‌جایی بر اساس اهمیت لغات موجود در متن ورودی، توانسته‌ایم کلمات کاندید را با کلمات موجود در نمونه‌های اصلی جابه‌جا کنیم تا جایی‌که نمونه مورد نظر در طرف دیگر مرز تصمیم طبقه‌بند قرارگیرد و مدل را در تشخیص دچار خطا کند. با این‌کار نیازی به جستجو در میان کلمات کاندید برای یافتن یک ترتیب مناسب از لغات برای جایگزینی نیست. در نتیجه این روش نسبت به برخی دیگر از روش‌ها سرعت اجرای بالایی نیز دارد.

روش پیشنهادی بدین صورت است که در ابتدا به‌ازای هر کدام از کلمات موجود در نمونه ورودی، از میان مجموعه کلمات موجود در لغت نامه، کلمه‌ای که بتواند میزان خطا در پیش‌بینی طبقه‌بند را افزایش‌دهد به‌عنوان کلمه کاندید برای جایگزینی انتخاب‌نموده و سپس با درنظرگرفتن اهمیت هر کلمه در مقدار هزینهٔ محاسبه‌شده کلمه کاندید مربوط، یک ترتیبی برای جایگزینی میان کلمات پیشنهاد می‌دهیم. در نهایت روش ارائه‌شده را روی دو مدل تحلیل احساس مختلف ارزیابی‌نموده و نشان‌داده‌ایم که این روش می‌تواند با دست‌کاری اندک نمونه‌های ورودی، مدل‌ها را وادار به اشتباه در تشخیص کند. آزمایش‌های مختلف روی روش ارائه‌شده نشان‌داده‌است که روش پیشنهادی در کاهش دقت هر دو مدل بسیار مؤثر بوده‌است به‌طوری‌که این روش توانسته دقت مدل را به کمتر از حدود ٪۱۰ کاهش‌دهد و این امر بیانگر موفقیت روش پیشنهادی نسبت به برخی دیگر از روش‌های مشابه است.

ساختار مقاله بدین‌صورت است که ابتدا در بخش دوم به شرح مفاهیم و تعاریف اولیه در رابطه با نمونه‌های تخاصمی پرداخته و سپس روش مورد استفاده در این پژوهش را در بخش سوم بیان‌می‌نماییم. در نهایت در بخش چهارم به تشریح نتایج حاصل و در بخش پنجم به جمع‌بندی مطالب پژوهش خواهیم‌پرداخت.

۲. بیان مسئله

طبقه‌بند شبکه عصبی f می‌تواند هر ورودی $x_i \in X = \{x_1, \dots, x_n\}$ را به برچسب $y_i \in Y = \{y_1, \dots, y_k\}$ با میزان احتمال بالایی نگاشت‌نماید. با افزودن اختلال بسیار کوچک روی نمونه‌های ورودی و دست‌کاری داده های ورودی می‌توان طبقه‌بند را موردحمله قرارداد تا دچار خطا شده و برچسبی غیر از برچسب اصلی را پیش‌بینی‌نماید. با این کار

می‌توان مدل‌های شبکه عصبی را بررسی و ارزیابی نموده و آن‌ها را نسبت به حملات تخاصمی مقاوم و آگاه نمود. میزان دستکاری روی داده‌های ورودی باید بسیار کم باشد از طرفی این تغییر باید بتواند دقت مدل را کاهش‌دهد. حال هرچقدر میزان تغییرات اعمال‌شده بر روی داده‌های ورودی به‌منظور گمراهی مدل کمتر باشند و بتوانند دقت مدل را به‌طرز مطلوبی کاهش‌دهند، حمله بهتری خواهیم‌داشت [۱۶].

۱،۲. حملات هدفمند

در حملات تخاصمی بدون هدف، با اعمال تغییرات اندک روی داده‌های ورودی مدل را وادارنموده تا برچسبی غیر از برچسب اصلی را پیش‌بینی‌نماید؛ درصورتی‌که در یک حمله هدفمند با دستکاری داده‌های ورودی، مدل را وادارمی‌نماییم تا یک برچسب مشخص و هدف، مانند y_{target} را پیش‌بینی‌نماید. اگر فرض‌کنیم نمونه تغییریافته به‌صورت x_{adv} باشد، در این حالت در جستجوی راهکاری برای حداقل‌سازی تفاوت میان پیش‌بینی مدل روی نمونه تغییر یافته و y_{target} با استفاده از دستکاری دادهٔ ورودی و در نتیجه محاسبه مسئله بهینه‌سازی زیر هستیم [۱۳]:

$$arg\min_{x_{adv}} loss(y_{target}, f(x_{adv})) \qquad (۱)$$

در رابطه ۱، $loss(y_{target}, f(x_{adv}))$ تابع خطای مدل را نمایش‌می‌دهد و این درواقع فاصله میان خروجی مدل روی نمونه تغییریافته شده و y_{target} را محاسبه‌می‌کند.

۳. روش پیشنهادی

در این بخش جزئیات روش پیشنهادی خود را به‌منظور تغییر داده‌های ورودی و تولید نمونه‌های تخاصمی ارائه‌خواهیم‌کرد و پس از تولید نمونه‌های تخاصمی، با تزریق آن‌ها به مدل شبکه عصبی، آن‌را دچار خطا در تشخیص می‌کنیم. هر متن ورودی از ترکیب تعدادی کلمه تشکیل‌می شود، بنابراین برای دستکاری ورودی در سطح کلمه، یک روش، جایگزینی بعضی از کلمات متن با کلمات مخربی است که بتوانند منجر به فریب مدل و در نتیجه تشخیص نادرست در خروجی مدل شوند. در ادامه به بیان چگونگی یافتن کلمات و جایگزینی با کلمات موجود در نمونه می‌پردازیم.

۱،۳. یافتن کلمات کاندید

در مرحله اول می‌بایست کلماتی را به‌عنوان کاندید برای جایگزینی با هرکدام از کلمات موجود در متن اصلی بیابیم. این کلمه کاندید باید به‌گونه‌ای باشد تا بعد از جایگزینی بتواند مدل را فریب‌دهد. برای این کار، به‌ازای هر کلمهٔ موجود در متن ورودی، از میان کلمات موجود در مجموعه لغت‌نامه، کلمه‌ای را به‌عنوان جایگزین انتخاب‌می‌کنیم که بیشترین مقدار هزینه را به مدل تحمیل‌نماید [۱۶،۱۷].

یک روش مناسب برای یافتن کلمات جایگزینی که بیشترین تأثیر منفی روی تصمیم طبقه‌بند دارند، بهره‌گیری از بسط تیلور مرتبه اول می‌باشد. با بهره‌گیری از بسط تیلور می‌توان سریع‌تر هزینه جابه‌جایی هرکدام از کلمات با کلمات موجود در لغت‌نامه را تخمین‌زد و به‌جای این‌که بعد از هر بار جابه‌جایی کلمات، هزینه دقیق را با درخواست‌زدن به مدل محاسبه‌نماییم، یک بار گرادیان تابع هزینه نسبت به مقادیر ورودی را طی یک فاز برگشت به عقب به‌دست‌می‌آوریم.

به عبارت دقیق‌تر در این حالت گرادیان هزینه مدل نسبت به هریک از کلمات ورودی طی یک فاز برگشت به عقب محاسبه‌شده و بردار خروجی حاصل، در ماتریس تعبیه برداری که در آن تمامی کلمات لغت‌نامه به صورت برداری نگاشت شده‌اند، ضرب‌می شود. در نتیجه با این کار برداری خواهیم‌داشت که در آن هزینه جابه‌جایی هر کلمه از متن اصلی با کلمات موجود در لغت‌نامه محاسبه‌شده و سپس با انتخاب بیشترین مقدار هزینه به‌ازای هر کلمه به‌همراه اندیس متناظر با آن، برداری به‌دست‌می‌آید که مناسب‌ترین جایگزین را به‌ازای هر لغت برای فریب مدل در خود دارد. رابطه ۲ این امر را نشان‌می‌دهد.

$$arg\max_{v_i \in V} [v_i]^T . \nabla_x loss \qquad (۲)$$

در این رابطه V بردار تعبیه تمامی کلمات موجود در لغت‌نامه است و $\nabla_x loss$ نیز بیانگر میانگین گرادیان خطای مدل روی نمونه‌های ورودی است.

۲،۳. جایگذاری کلمات

در مسائل طبقه‌بندی متنی، هر کلمه موجود در متن، دارای ارزش و تأثیری در تصمیم‌گیری طبقه‌بند است. به‌عنوان مثال در عبارت « such a good movie» کلمه good نسبت به سایر کلمات اهمیت بیشتری دارد. در نتیجه طبقه‌بند با ارزیابی آن عبارت، کل متن را به‌صورت یک متن مثبت پیش‌بینی‌می‌کند. حال بعد از آن‌که به‌ازای هریک از کلمات موجود در متن اصلی، کلمه کاندیدی برای جایگزینی یافت‌شد، در مرحله بعدی برای مشخص‌نمودن اولویت جابه‌جایی هر کدام از کلمات با کلمات کاندید خود، ارزش تمامی کلمات موجود در متن اصلی را به‌صورت جداگانه محاسبه‌نموده و به‌صورت یک بردار وزنی در بردار هزینه جابه‌جایی کلمه مورد نظر که در مرحله قبل محاسبه‌نمودیم، ضربمی‌کنیم. در نهایت نیز مقادیر به‌دست‌آمده را به‌ترتیب بزرگترین مقادیر مرتبمی‌کنیم و با این کار به هر کلمه‌ای برای جایگزینی با کلمه متناظر خود امتیازمی‌دهیم تا کلمات بر اساس امتیاز کسب شده خود جایگزین شوند تا زمانی‌که مدل دچار خطا شود.

ارزش هر کلمه را که با $S(x, w_i)$ نمایش‌داده‌می‌شود، می‌توان از طریق محاسبه اختلاف احتمال وقوع برچسب واقعی روی نمونه ورودی و احتمال وقوع برچسب واقعی در صورت حذف کلمه مذکور و قراردادن توکن <UNK> در آن جایگاه، به‌صورت رابطه زیر تعریف‌کرد [۱۵]:

$$S(x, w_i) = p(y_{true}|x) - p(y_{true}|\hat{x}_i) \qquad (۳)$$

به‌طوری‌که:

$$x = w_1, \dots, w_i, \dots w_d$$

$$\hat{x}_i = w_1, \dots, < UNK >, \dots, w_d$$

در رابطه (۳)، x، متن ورودی به‌طول d کلمه و $\hat{x}_i$ متن ورودی است که در آن توکن <UNK> جایگزین کلمه w_i شده‌است. به همین صورت، ارزش همه کلمات موجود در متن، با استفاده از رابطه ۳ محاسبه‌می‌گردد. در نتیجه به‌ازای متن x یک بردار $S(x)$خواهیمداشت. در ادامه، مطابق با رابطه ۴، بردار نهایی را در بردار هزینه کلماتی که در مرحله قبل محاسبه‌شده، ضرب نموده و خروجی نهایی را به‌ترتیب نزولی مرتب‌می‌کنیم. این ترتیب به‌دست آمده نشان‌دهنده ترتیب جایگذاری هریک از کلمات کاندید با کلمه متناظر خود در متن اصلی است. درواقع با این‌کار کلمات را براساس اهمیت و بیشترین میزان خطا جایگذاری‌نموده‌ایم. در انتها نیز پس از هر جایگذاری یک کلمه با کلمه کاندید خود به‌ترتیب اولویت یافت‌شده، خروجی مدل محاسبه‌می‌شود و جابه‌جایی کلمات تا زمانی که پیش‌بینی مدل تغییرکند، ادامه‌می‌یابد. این رویکرد نسبت به روش‌های جستجو برای جابه‌جایی کلمات مناسب، سریع‌تر است. رابطه ۴ تابع محاسبه اولویت جابه‌جایی کلمات را نشان‌می‌دهد.

$$H(x, x_i^*, w_i) = \phi(S(x))_i . \phi([v_i]^T . \nabla_x loss)_i \qquad (۴)$$

در این رابطه $\phi(z)_i$ بیانگر تابع سافتمکس روی هر کدام از بردار مقادیر گرادیان و ارزش کلمات است. از این تابع برای نرمال‌سازی هر کدام از مقادیر ارزش و هزینه، بین صفر و یک استفاده‌می‌نماییم که مطابق با رابطه (۵) محاسبه‌می‌شود.

$$\phi(z)_i = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}} \qquad (۵)$$

۴. نتایج آزمایش‌ها

مجموعه داده گان. در این پژوهش از پیکره با نک درختی تحلیل احساس داشگاه استنفورد استفاده موده‌ایم [۱۹]. این پیکره شامل مجموعه‌ای از نظرات بینندگان فیلم به زبان انگلیسی است که به‌صورت دوکلاسه و پنج‌کلاسه آن موجود است و در این آزمایش از نوع دوکلاسه این پیکره (SST-2) به‌صورت نظرات مثبت با برچسب ۰ و نظرات منفی با برچسب ۱ استفاده‌نموده‌ایم.

شبکه‌های قربانی. در این پژوهش از دو مدل شبکه عصبی به‌عنوان شبکه‌های قربانی برای اجرای الگوریتم روی آن‌ها به‌صورت زیر استفاده‌کرده‌ایم:

مدل حافظه طولانی کوتاه‌مدت با دو لایه پنهان (LSTM) که اغلب از این مدل برای مسائل طبقه‌بندی متون استفاده‌می شود. در این مدل لایه آخر یک لایه کاملاً متصل است که خروجی لایه پنهان را دریافت‌نموده و پیش‌بینی نهایی را محاسبه‌می‌نماید.

مدل شبکه عصبی پیچشی با دو لایه پنهان (CNN) و لایه آخر این مدل نیز یک لایه کاملا متصل برای محاسبه کلاس نهایی استفاده‌شده‌است.

در هر دو مدل از یک لایه تعبیه برای نگاشت نمونه‌های ورودی به فضای پیوسته و بردار تعبیه کلمات word2vec با فضای تعبیه ۳۰۰ بعدی استفاده‌نموده‌ایم [۲۰].

۱،۴. ارزیابی

الگوریتم حمله پیشنهادی در این پژوهش را پس از پیاده‌سازی، با الگوریتم PWWS، به‌عنوان یک روش پایه مقایسه‌نموده‌ایم [۱۵]. رویکرد ارائه‌شده در PWWS یکی از روش‌هایی است که توانسته نسبت به سایر روش‌های ارائه‌شده در حوزه تولید نمونه‌های تخاصمی، با میانگین نرخ جابه جایی و تغییرات اندک روی متن ورودی، موفق عمل‌نماید. بنابراین در پژوهش‌های مختلف، به‌عنوان یک روش پایه برای سنجش و مقایسه استفاده‌می‌شود. از طرفی این الگوریتم با دست‌کاری و جابه‌جایی نمونه‌های ورودی موفق‌شده به‌طرز مطلوبی دقت مدل را روی برخی از پیکره‌های ارائه‌شده برای طبقه‌بندی متون نیز کاهش‌دهد.

روش حمله PWWS به‌صورت جایگزینی کلمات به‌طرز حریصانه است به‌طوری‌که در آن با محاسبه اهمیت هریک از کلمات موجود در جمله ورودی و هم‌چنین میزان تأثیر جابه‌جایی هرکدام از کلمات با مترادف خود آن هم از طریق محاسبه اختلاف خروجی مدل روی نمونه اصلی با خروجی مدل پس از جایگزینی کلمه مورد نظر با کلمه مترادف خود، رویکردی برای جایگزینی لغات ارائه‌شده‌است. در این روش به جای انتخاب کلمات جایگزین از فضای برداری، هر کلمه با کلمه مترادف خود از مجموعه wordnet جایگزین‌می‌شود و در انتها کلمه‌ای که بیشترین مقدار خطا را در پیش‌بینی طبقه‌بند دارد، به‌عنوان کلمه مناسب برای جایگزینی، انتخاب‌می‌گردد.

برای سادگی در پیاده‌سازی این روش، موجودیت‌های نامدار را نادیده‌گرفته و برای مقایسه، هر دوی حملات را به‌صورت حمله هدفمند تعریف‌کرده‌ایم. به‌طوری‌که مدل به‌ازای هر نمونه‌ای برچسب مخالف با برچسب اصلی خود را پیش‌بینی‌نماید. نتایج حاصل از پیاده‌سازی و مقایسه دو روش در جدول‌های ۱و۲ قراردارند.

جدول ۱: دقت مدل قبل و بعد از حمله

کلاس داده‌های تست	مدل	دقت مدل روی داده‌های هدف قبل از حمله	دقت مدل با روش PWWS	دقت مدل با روش پیشنهادی
منفی (۱)	LSTM	۸۵,۲۸٪	۴۷,۶۶٪	**۱,۶۳٪**
مثبت (۰)	LSTM	۸۸,۲۸٪	۴۳,۹۱٪	**۰,۰۵٪**
منفی (۱)	CNN	۷۷,۳۳٪	۵۱,۶۳٪	**۱۰,۰۴٪**
مثبت (۰)	CNN	۸۶,۰۳٪	۷۷,۲۵٪	**۷,۲٪**

به‌منظور ارزیابی الگوریتم پیشنهادی و بررسی کیفیت نمونه‌های متنی تولیدشده و مقایسه دو روش حمله، در ابتدا پس از پاکسازی و پیش‌پردازش اولیه، تمامی نمونه‌های موجود در مجموعه داده‌های تست با برچسب منفی و مثبت را از یکدیگر جدانمودیم تا بتوانیم به صورت جداگانه روی هریک الگوریتم حمله هدفمند را اجراکنیم به‌گونه‌ای که مدل، برچسب هدف و مخالف با برچسب اصلی را پیش‌بینی‌کند (برای داده‌های منفی برچسب هدف را ۰ و برای داده‌های مثبت برچسب هدف را ۱ درنظرگرفته‌ایم).

در مرحله بعدی، مدل ازپیش‌آموزش‌دیده خود را بررسی‌نموده و قبل از اعمال حمله، دقت مدل را به‌صورت جداگانه روی هریک از داده-های با برچسب مثبت و منفی اندازه‌گیری‌کردیم. مطابق با نتایج جدول ۱ دقت مدل LSTM روی مجموعه داده‌های با کلاس مثبت در حدود ٪۸۸ و برای داده‌های منفی در حدود ٪۸۵ شده‌است. دقت مدل CNN نیز روی داده‌های با کلاس مثبت حدود ٪۸۶ و برای داده‌های منفی حدود ٪۷۷ گزارش‌شده‌است. حال در انتها پس از دست‌کاری و تغییر نمونه‌های موجود در مجموعه داده‌ها و اعمال الگوریتم حمله روی هرکدام از مدل‌ها می‌بایست با محاسبه دقت مدل روی داده‌های تغییریافته میزان گمراهی مدل‌هایمان را مشخص‌نماییم. بعد از اجرای الگوریتم حمله، هرچقدر دقت مدل در تشخیص درست کمتر باشد، حمله قوی‌تری خواهیم‌داشت. با مقایسه دو روش، دیده‌می‌شود که روش پیشنهادی در کاهش دقت بسیار خوب عمل‌کرده‌است و توانسته دقت مدل LSTM را روی هر دو نوع داده مثبت و منفی، به زیر ٪۲ برساند و برای مدل CNN نیز دقت به زیر ٪۱۰ رسیده‌است. با توجه به نتایج آمده در جدول۱، دقت مدل LSTM روی داده‌های منفی با روش پیشنهادی به ٪۱,۶۳ و روی داده‌های مثبت ٪۰,۰۵ رسیده‌است درحالی‌که دقت این مدل با روش PWWS روی دو داده مثبت و منفی به حدود ٪۴۰ رسیده-است همین‌طور روی مدل CNN نیز دقت روی داده‌های منفی پس از حمله به حدود ٪۱۰ و برای داده‌های مثبت به حدود ٪۷ رسیده‌است که در مقایسه با الگوریتم PWWS به‌شدت افت دقت داشته‌ایم که این امر، نشان‌دهنده مؤثربودن روش حمله پیشنهادی است.

برای این‌که حمله تخاصمی مناسب باشد، باید تغییرات اعمال‌شده روی نمونه‌های ورودی بسیار اندک باشند تا از دید انسان این میزان تغییرات قابل‌تشخیص نباشد. روش پیشنهادی در این پژوهش توانسته با جابه‌جایی و تغییر تعداد کمی از کلمات نیز در گمراهی هریک از مدل‌ها و کاهش دقت، بسیار موفق عمل‌نماید.

جدول ۲ میانگین نرخ جایگزینی کلمات پس از اجرای حمله هدفمند روی هرکدام از داده‌های با برچسب مثبت و منفی را نشان‌می دهد. میزان نرخ جابه‌جایی کلمات برابر با تعداد جایگزینی‌های کلمات در یک نمونه بر کل تعداد کلمات موجود در نمونه مورد نظر است. هرچقدر میانگین نرخ جابه‌جایی کلمات با اعمال الگوریتم حمله پایین‌تر باشد ،بهتر است چراکه به‌نوعی نامحسوس بودن این تغییرات از دید انسان را نشان‌می‌دهد. در روش پیشنهادی میانگین جابه‌جایی‌ها برای فریب مدل LSTM روی داده‌های منفی و مثبت به‌ترتیب ٪۱۳,۹۸ و ٪۱۰,۸۹ شده‌است، درحالی‌که در روش PWWS این میزان حدود ٪۱۹و٪۱۷گزارش‌شده‌است و در مدل CNN نیز میانگین نرخ جابه‌جایی روی داده‌های منفی و مثبت به‌ترتیب ٪۹,۴۵ و٪۸,۱۵ شده‌است، درحالی‌که در روش PWWS این میزان حدود ٪۲۵ و٪۱۷گزارش‌شده-است. روش پیشنهادی در این مورد نیز روی هر دو نوع مدل شبکه عصبی نسبت به روش PWWS بهتر عمل‌کرده‌است.

جدول ۲: میانگین نرخ جابجایی کلمات پس از حمله

میانگین نرخ جابجایی کلمات در روش پیشنهادی	میانگین نرخ جاجایی کلمات در روش PWWS	مدل	کلاس مجموعه داده تست
۱۳,۹۸٪	۱۹,۲۲٪	LSTM	منفی(۱)
۱۰,۸۹٪	۱۷,۸۴٪	LSTM	مثبت(۰)
۹,۴۵٪	۱۷,۸۳٪	CNN	منفی(۱)
۸,۱۵٪	۲۵,۴۸٪	CNN	مثبت(۰)

برای مقایسه نمونه‌های تولیدی و تغییریافته با نمونه‌های اصلی و تصویرسازی بهتر، در جدول ۳، تعدادی از نمونه‌های تولیدشده با استفاده از روش پیشنهادی و روش PWWS را نشان‌داده‌ایم و خروجی مورد نظر را روی مدل LSTM اندازه‌گیری‌نموده‌ایم. به‌ازای هر کدام از حملات هدفمند و دست‌کاری نمونه‌های موجود در داده‌های منفی و مثبت، یک نمونه اصلی را همراه نمونه تغییریافته‌اش قرارداده‌ایم. در هر یک از مثال-ها پیش‌بینی مدل LSTM، قبل و بعد از اجرای الگوریتم حمله و نیز میزان درجه اطمینان برای هر پیش‌بینی نیز اندازه‌گیریشده‌است. کلمات تغییریافته در جدول با رنگ قرمز مشخص‌شده‌اند.

از بررسی و مقایسه موارد موجود در جدول می‌توان مشاهده‌کرد که میزان تغییرات صورت‌گرفته روی نمونه‌های ا صلی با روش پی شنهادی نسبت به روش PWWS بسیار کمتر است. مثلاً در روش PWWS نمونه اول جدول که توسط مدل به‌صورت منفی پیش‌بینی‌شده، با چهار تغییر از کلاس منفی به کلاس مثبت انتقال‌یافته و با درجه اطمینان ٪۶۰,۶۸ مدل را گمراه‌کرده‌است. درحالی‌که در روش پیشنهادی، همان نمونه تنها با یک تغییر توانسته با درجه اطمینان نسبتاً بالایی (٪۹۲,۵۵) مرز تصمیم مدل LSTM را جابه‌جا کرده و منجر به تشخیص خطا شود. برای یک نمونه مثبت نیز می‌توان مشاهده‌کرد که در روش PWWS، مدل با درجه اطمینان کمتر(٪۵۳,۲) و تغییرات بیشتری (سه تغییر) دچار خطا شده‌است. درحالی‌که روش پیشنهادی تنها با یک تغییر و درجه اطمینان بیشتر از ٪۹۵ موفق به گمراهی مدل شده‌است.

جدول ۳: مثالی از نمونه‌های تولیدشده

نمونه تغییریافته	نمونه اصلی	پیش‌بینی بعد از حمله	پیش‌بینی قبل از حمله	روش حمله
zaidan's script has just decent plot to string the stunts together and not rather enough word to keep the faces straight.	zaidan's script has barely enough plot to string the stunts together and not quite enough characterization to keep the faces straight.	مثبت (۰) درجه اطمینان ٪۶۰,۸۶	منفی (۱) درجه اطمینان ٪۹۸,۳۱	روش PWWS
zaidan's script has barely enough plot to string the stunts together and infectiously quite enough characterization to keep the faces straight.	zaidan's script has barely enough plot to string the stunts together and not quite enough characterization to keep the faces straight.	مثبت (۰) درجه اطمینان ٪۹۲,۵۵	منفی (۱) درجه اطمینان ٪۹۸,۳۱	روش پیشنهادی
Seldom has a movie therefore nearly matched the tone of a man and his work.	Seldom has a movie so closely matched the spirit of a man and his work.	منفی (۱) درجه اطمینان ٪۵۳,۲۰	مثبت (۰) درجه اطمینان ٪۹۶,۲۰	روش PWWS
ill-constructed has a movie so closely matched the spirit of a man and his work.	Seldom has a movie so closely matched the spirit of a man and his work.	منفی (۱) درجه اطمینان ٪۹۵,۴۰	مثبت (۰) درجه اطمینان ٪۹۶,۲۰	روش پیشنهادی

با توجه به نتایج آزمایش‌ها می‌توان مشاهده‌کرد که روش پیشنهادی در تمام موارد بررسی‌شده از جمله کاهش میزان دقت مدل، ایجاد تغییرات بسیار کم و تغییر مرز تصمیم طبقه‌بند با درجه اطمینان بالا، توانسته بسیار موفق عمل‌نماید، به‌طوری‌که مقایسه‌های مختلف با روش PWWS بیانگر مؤثربودن این روش در فریب مدل‌ها و یافتن نقاط ضعف و آسیب‌پذیری‌های شبکه عصبی نسبت به این قبیل از نمونه‌ها است. بنابراین می‌توان از این رویکرد برای تولید نمونه‌های متنی، مشابه با نمونه‌های اولیه و با کمترین میزان تغییر، برای بررسی آسیب‌پذیری‌های مدل و در انتها مقاوم سازی مدل‌های شبکه عصبی نسبت به این قبیل از نمونه‌ها استفاده‌کرد.

۵. نتیجه‌گیری و کارهای آتی

به‌دلیل گسسته‌بودن ماهیت متن، ساخت نمونه‌های تخاصمی متنی امری چالش‌برانگیز است. تغییرات اعمال‌شده روی نمونه‌های متنی نیز باید اندک باشند تا از دید انسان پنهان‌مانده و به‌سادگی قابل‌شناسایی و تشخیص نباشند. در این پژوهش، روشی بهینه برای تولید نمونه‌های تخاصمی متنی به‌منظور حمله روی مدل‌های یادگیری عمیق ارائه‌شد. که در آن برای یافتن کلمات جایگزین، به‌ازای هرکدام از کلمات موجود در متن اصلی، کلمه‌ای که بیشترین هزینه را در هر جایگزینی دارد با استفاده از گرادیان تابع هزینه، از میان لغات موجود در لغت نامه انتخاب‌نموده و با درنظرگرفتن بیشترین اهمیت هر لغت همراه میزان هزینه محاسبه‌شده برای جایگذاری آن‌ها اولویتی برای جایگزینی کلمات ارائه‌شد.

نتایج آزمایش‌ها نشان‌داده‌اند که با اجرای الگوریتم پیشنهادی روی نمونه‌های ورودی و دست‌کاری نمونه‌ها و تزریق نمونه‌های تغییریافته به مدل، هریک از مدل‌ها دچار افت دقت شدیدی شدند طوری‌که این میزان برای مدل LSTM به کمتر از ٪۲ و برای یک مدل CNN به کمتر از ٪۱۰ رسیده‌است. از سوی دیگر، با محاسبه میانگین نرخ جابجایی کلمات نشان‌داده‌شد که با استفاده از روش پیشنهادی می‌توان تنها با جایگزینی تعداد کمی از کلمات موجود در متن ورودی با کلماتی دیگر، مدل را به‌اشتباه‌انداخته و وادار به تشخیص نادرست نمود.

در آینده تلاش خواهیم‌کرد تمرکز بیشتری روی محتوای متون تولیدی داشته‌باشیم و کلمات موجود در متن را با کلماتی متناسب و مرتبط با متن ورودی جایگزین نماییم طوری‌که علاوه بر پایین نگه‌داشتن دقت مدل و اعمال تغییرات اندک، متون تولید شده از لحاظ مفهومی نیز قابل‌تشخیص توسط انسان را نباشند و با ارائه روش‌های مختلف دفاعی بتوانیم مدل‌های شبکه عصبی را نسبت به این قبیل حملات آگاه و مقاوم‌سازیم. از سویی دیگر، نمونه‌های تخاصمی تولیدشده علاوه بر بحث مقاوم‌سازی مزایای جانبی زیادی نیز دارند یکی از این موارد تقویت دادگان می‌باشد، بنابراین به‌نظر می‌رسد با بهبود روش‌های تولید نمونه‌های تخاصمی در آینده بتوانیم از این قبیل نمونه‌ها در جهت تقویت دادگان ناکافی نیز استفاده‌نماییم.

مراجع

[1] C.Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhun, I. Goodfellow and R. Fergus, "Intriguing properties of neural networks", 2nd International Conference on Learning Representations, ICLR 2014, Banff, Canada, 2014.

[2] R. Jia., P. Liang, Adversarial examples for evaluating reading comprehension systems. In EMNLP, 2017

[3] Y. Belinkov and Y. Bisk, "Synthetic and natural noise both break neural machine translation", In Proceedings of ICLR, 2018.

[4] I. Fursov, A. Zaytsev, P. Burnyshev, E. Dmitrieva, N. Klyuchnikov, A. Kravchenko, E. aArtemova and E. Burnaev, "A differentiable language model adversarial attack on text classifiers", arXiv:2107.11275v1 [cs.CL], 23 Jul 2021.

[5] Z. Kong, J. Xue, Y. Wang, L. Huang, Z. Niu and E. Li, "A survey on adversarial attack in the age of artificial intelligence", Wireless Communications and Mobile Computing, Volume 2021, Article ID 4907754, 22 pages, 2021.

[6] J. Xu and Q. Du, "TextTricker:Loss-based and gradient-based adversarial attacks on text classification models", Engineering Applications of Artificial Intelligence,Volume 92, Elsevier, 0952-1976, 2020.

[7] H. Hosseini, S. Kannan, B. Zhang and R. Poovendran, “Deceiving google’s perspective api built for detecting toxic comments,” arXiv preprint arXiv:1702.08138, 2017.

[8] M. Alzantot, Y. Sharma, A. Elgohary, B. Ho, M. Srivastava and K. Chang, “Generating natural language adversarial examples, in Proceedings of Conference on Empiritical Methods in Natural Language Processing (EMNLP), 2018.

[9] B. Liang, H. Li, M. Su, P. Bian, X. Li and W. ChangShi, “Deep text classification can be fooled”, arXiv preprint arXiv:1704.08006, 2017.

[10] S. Samanta and S. Mehta, “Towards crafting text adversarial samples”, arXiv preprint arXiv:2003.10388, July 2017.

[11] N. Papernot, P. McDaniel, A. Swami and R. Harang,“Crafting adversarial input sequences for recurrent neural networks”, In 2016 IEEE Military Communications Conference, MILCOM 2016, Baltimore, MD, USA, p.p. 49–54, November 1-3, 2016.

[12] M. Sato, J. Suzuki, H. Shindo and Y. Matsumoto, “Interpretable adversarial perturbation in input embedding space for text”, In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, (IJCAI 2018), Stockholm, Sweden, p.p. 4323– 4330, July 13-19, 2018.

[13] M. Behjati, S. M. Moosavi-Dezfooli, M. SoleymaniBaghshah and P. Frossard, “Universal adversarial attacks on text classifiers”, In ICASSP, 2019.

[14] L. Song, X. Yu, H. Peng and K. Narasimhan, “Universal adversarial attacks with natural triggers for text classification”, arXiv:2005.00174v2 [cs.CL], 7 Apr 2021.

[15] S. Ren, Y. Deng, H. He and W. Che,“Generating natural language adversarial examples through probability weighted word saliency”, In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, p.p. 1085–1097, 2019.

[16] J. Ebrahimi, A. Rao, D. Lowd and D. Dou, “Hotflip: White-box adversarial examples for text classification”, In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers, p.p. 31–36, 2018.

[17] E. Wallace, S Feng, N. Kandpal, M. Gardner and S. Singh, “Universal adversarial triggers for attacking and analyzing nlp”, arXiv preprint arXiv:1908.07125, 2019.

[18] H. Zhang, H. Zhou, N. Miao and L. Li, “Generating fluent adversarial examples for natural languages”, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019.

[19] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D Manning, A. Ng and C. Potts, “Recursive deep models for semantic compositionality over a sentiment treebank”, in Proceedings of the conference on empirical methods in natural language processing (EMNLP), p.p. 1631–1642, 2013.

[20] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch and A. Joulin, “Advances in pre-training distributed word representations”, In LREC, 2018.