تشخیص کلاهبرداری مخابراتی مبتنی بر متن با استفاده از مدل‌های Transformer : مطالعه چندزبانه بر پایهٔ CHIFRAUD

۱ . مقدمه

با گسترش ارتباطات دیجیتال و افزایش وابستگی کاربران به پیامک‌ها و پیام‌های متنی، کلاهبرداری مخابراتی به یکی از چالش‌های جدی امنیت اطلاعات و حریم خصوصی تبدیل شده است. پیام‌های جعلی، تبلیغات فریبنده، و تلاش برای سرقت اطلاعات مالی یا هویتی کاربران، نه‌تنها خسارات اقتصادی قابل توجهی ایجاد می‌کنند، بلکه اعتماد عمومی به زیرساخت‌های ارتباطی را نیز تضعیف می‌نمایند. تشخیص دستی چنین پیام‌هایی به دلیل حجم بالای داده و پیچیدگی زبانی، عملاً غیرممکن بوده و نیازمند راهکارهای خودکار و مقیاس‌پذیر است.

در سال‌های اخیر، روش‌های پردازش زبان طبیعی (Natural Language Processing) نقش مهمی در شناسایی الگوهای زبانی مرتبط با اسپم و کلاهبرداری ایفا کرده‌اند. با این حال، بسیاری از مطالعات پیشین محدود به زبان انگلیسی بوده و قابلیت تعمیم به زبان‌های دیگر را نداشته‌اند. این مسئله به‌ویژه در حوزهٔ کلاهبرداری مخابراتی، که الگوهای زبانی آن در فرهنگ‌ها و زبان‌های مختلف تفاوت چشمگیری دارند، اهمیت بیشتری پیدا می‌کند.

در این پروژه، یک چارچوب چندزبانه برای تشخیص کلاهبرداری متنی ارائه شده است که سه زبان چینی، انگلیسی و عربی را پوشش می‌دهد. هستهٔ اصلی این مطالعه مبتنی بر دیتاست CHIFRAUD است که به‌عنوان یک بنچمارک بلندمدت برای تشخیص کلاهبرداری مخابراتی در زبان چینی معرفی شده است. علاوه بر پیاده‌سازی نسخهٔ اصلی مقاله برای زبان چینی، دو پیاده‌سازی تکمیلی برای زبان‌های انگلیسی و عربی نیز ارائه شده‌اند تا قابلیت تعمیم روش به زبان‌های دیگر بررسی شود. تمرکز اصلی این گزارش بر تحلیل روش، تنظیمات آزمایشگاهی و ارزیابی دقیق نتایج به‌دست‌آمده از این پیاده‌سازی‌ها است.

---

۲ . کارهای مرتبط

روش‌های اولیهٔ تشخیص اسپم عمدتاً بر پایهٔ ویژگی‌های سطحی متن و الگوریتم‌های یادگیری ماشین کلاسیک مانند Naive Bayes، Support Vector Machines و Logistic Regression بنا شده بودند. این روش‌ها معمولاً نیازمند استخراج دستی ویژگی‌ها بوده و در مواجهه با تغییر الگوهای زبانی یا داده‌های نویزی عملکرد محدودی داشتند. همچنین، تعمیم این روش‌ها به زبان‌های مختلف اغلب مستلزم طراحی مجدد ویژگی‌ها بود.

با ظهور یادگیری عمیق، مدل‌هایی مانند شبکه‌های عصبی بازگشتی و کانولوشنی وارد حوزهٔ تشخیص اسپم شدند و توانستند وابستگی‌های زبانی پیچیده‌تری را مدل‌سازی کنند. با این حال، این مدل‌ها همچنان در درک وابستگی‌های بلندمدت و انتقال دانش بین زبان‌ها محدودیت‌هایی داشتند.

معرفی معماری Transformer و مدل‌های از پیش آموزش‌دیده‌ای مانند BERT نقطهٔ عطفی در پردازش زبان طبیعی ایجاد کرد. این مدل‌ها با استفاده از مکانیزم self-attention قادر به استخراج نمایش‌های معنایی عمیق از متن هستند و در بسیاری از وظایف طبقه‌بندی متنی عملکردی فراتر از روش‌های پیشین ارائه داده‌اند. در حوزهٔ تشخیص اسپم و کلاهبرداری، استفاده ازTransformer ها امکان شناسایی الگوهای ظریف زبانی و تطبیق بهتر با داده‌های نامتوازن را فراهم کرده است.

دیتاست CHIFRAUD در این میان جایگاه ویژه‌ای دارد، زیرا برخلاف بسیاری از مجموعه‌داده‌های پیشین که صرفاً دودسته‌ای هستند، یک مسئلهٔ چندکلاسهٔ واقعی را در حوزهٔ کلاهبرداری مخابراتی مدل می‌کند و به‌عنوان یک معیار استاندارد برای ارزیابی روش‌های جدید مطرح شده است.

---

۳ . داده‌ها(Datasets)

در این پروژه از سه دیتاست مستقل برای سه زبان مختلف استفاده شده است که هر یک ویژگی‌ها و چالش‌های خاص خود را دارند.

دیتاست CHIFRAUD شامل پیام‌های متنی زبان چینی با برچسب‌های چندکلاسه است. این دیتاست دارای ۱۰ کلاس مختلف بوده و توزیع برچسب‌ها به‌شدت نامتوازن است؛ به‌طوری که کلاس صفر (پیام‌های عادی) بخش عمده‌ای از داده‌ها را تشکیل می‌دهد و برخی کلاس‌ها تنها تعداد محدودی نمونه دارند. این ویژگی، CHIFRAUDرا به یک مسئلهٔ چالش‌برانگیز و واقع‌گرایانه برای تشخیص کلاهبرداری تبدیل می‌کند.

برای زبان انگلیسی، از دیتاست SMS Spam Collection استفاده شده است که شامل پیام‌های کوتاه با دو برچسب hamو spam است. این دیتاست اگرچه نسبتاً کوچک‌تر است، اما به‌عنوان یک معیار کلاسیک در ارزیابی الگوریتم‌های تشخیص اسپم شناخته می‌شود. عدم توازن کلاسی در این مجموعه نیز وجود دارد، اما شدت آن نسبت به دیتاست چینی کمتر است.

دیتاست Arabic Spam Tweets Dataset برای زبان عربی به‌کار گرفته شده است. این مجموعه از داده‌های استخراج‌شده از Twitter تشکیل شده و شامل پیام‌های اسپم و غیر اسپم است. ویژگی مهم این دیتاست، تنوع زبانی و غیررسمی بودن متن‌ها است که چالش‌های خاصی را برای مدل‌سازی ایجاد می‌کند.

در تمامی دیتاست‌ها، ملاحظات مربوط به لایسنس و استفادهٔ اخلاقی رعایت شده و داده‌های خام در مخزن پروژه ذخیره نشده‌اند. تنها لینک دسترسی و مراحل پیش‌پردازش ارائه شده است.

---

۴ . روش پیشنهادی(Methodology)

چارچوب کلی سیستم در هر سه زبان ساختاری مشابه دارد. ابتدا داده‌ها بارگذاری شده و پس از پاک‌سازی اولیه، به مجموعه‌های آموزش و اعتبارسنجی با استفاده از تقسیم‌بندی stratified تقسیم می‌شوند تا توزیع برچسب‌ها حفظ شود. سپس متن‌ها با استفاده از tokenizer مدل مربوطه به نمایش‌های عددی تبدیل می‌شوند.

برای زبان چینی، مدل bert-base-chinese به‌عنوان مدل پایه انتخاب شده است. در زبان انگلیسی از distilroberta-baseاستفاده شده که نسخه‌ای سبک‌تر از RoBERTa با کارایی مناسب است. برای زبان عربی نیز مدل asafaya/bert-base-arabic به‌کار گرفته شده که به‌طور خاص برای متون عربی آموزش دیده است.

طول توالی ورودی برای هر زبان بر اساس ماهیت داده‌ها تنظیم شده است. به‌منظور مدیریت عدم توازن کلاسی، از تابع هزینهٔ weighted cross-entropy استفاده شده است که وزن هر کلاس بر اساس فراوانی آن در داده‌های آموزشی محاسبه می‌شود. این رویکرد باعث می‌شود مدل به کلاس‌های کم‌نمونه توجه بیشتری نشان دهد.

برای بهبود پایداری آموزش، از تکنیک‌های gradient accumulation و gradient clipping استفاده شده است. همچنین، زمان‌بندی نرخ یادگیری به‌صورت warmup به‌همراه کاهش خطی (linear decay) پیاده‌سازی شده است تا از نوسانات شدید در ابتدای آموزش جلوگیری شود.

---

۵ . تنظیمات آزمایش‌ها

در تمامی آزمایش‌ها، داده‌ها به‌صورت stratified به مجموعه‌های آموزش و اعتبارسنجی تقسیم شده‌اند. به‌منظور کاهش هزینهٔ محاسباتی و ایجاد تعادل نسبی بین کلاس‌ها، از راهبرد class capping استفاده شده است؛ به‌طوری که حداکثر تعداد نمونه برای هر کلاس محدود شده است.

هایپرپارامترهایی مانندbatch size ، تعدادepoch ها، نرخ یادگیری و weight decay بر اساس تنظیمات رایج در آموزش مدل‌های Transformer انتخاب شده‌اند. تمامی آموزش‌ها روی CPU انجام شده است که نشان‌دهندهٔ قابلیت اجرای مدل‌ها در شرایط سخت‌افزاری محدود نیز می‌باشد.

---

۶ . نتایج

برای ارزیابی عملکرد مدل‌ها از معیارهایAccuracy ، Macro-F1و Weighted-F1 استفاده شده است. در داده‌های نامتوازن، Macro-F1اهمیت ویژه‌ای دارد زیرا عملکرد مدل را به‌صورت مستقل از توزیع کلاس‌ها ارزیابی می‌کند.

در مدل چینی مبتنی برCHIFRAUD ، عملکرد بسیار بالایی در اعتبارسنجی مشاهده شده است و مقدار Macro-F1 نزدیک به 0.95 به‌دست آمده است که نشان‌دهندهٔ توانایی مدل در تشخیص کلاس‌های کم‌نمونه است. در مدل انگلیسی، نتایج دودسته‌ای با Macro-F1 حدود 0.98 حاصل شده که عملکردی بسیار مطلوب محسوب می‌شود. مدل عربی نیز با وجود چالش‌های زبانی، به Macro-F1 نزدیک به 0.99 دست یافته است.

تحلیل کیفی ماتریس‌های درهم‌ریختگی نشان می‌دهد که بیشتر خطاها مربوط به کلاس‌هایی با شباهت معنایی بالا هستند و مدل در تشخیص پیام‌های کاملاً جعلی عملکرد بسیار دقیقی دارد.

---

۷ . تحلیل کیفی

آزمایش‌های دستی بر روی نمونه‌های متنی نشان می‌دهد که مدل‌ها قادر به تشخیص الگوهای رایج کلاهبرداری مانند درخواست اطلاعات حساس، وعدهٔ جوایز مالی و لینک‌های مشکوک هستند. با این حال، در برخی موارد پیام‌های خنثی با لحن رسمی ممکن است به اشتباه به‌عنوان اسپم شناسایی شوند، که این موضوع به همپوشانی زبانی بین پیام‌های واقعی و جعلی بازمی‌گردد.

تفاوت رفتار مدل‌ها در زبان‌های مختلف نیز قابل توجه است. در زبان چینی، مسئلهٔ چندکلاسه چالش‌برانگیزتر بوده، در حالی که در زبان‌های انگلیسی و عربی تمرکز اصلی بر تشخیص دودسته‌ای است.

---

۸ . بحث

نتایج به‌دست‌آمده نشان می‌دهد که مدل‌های Transformer به‌دلیل توانایی در استخراج نمایش‌های معنایی عمیق، ابزار مناسبی برای تشخیص کلاهبرداری متنی هستند. تفاوت بین تشخیص دودسته‌ای و چندکلاسه اهمیت طراحی دقیق تابع هزینه و معیارهای ارزیابی را برجسته می‌کند.

از جمله محدودیت‌های این مطالعه می‌توان به اندازهٔ محدود برخی دیتاست‌ها، آموزش روی CPU و عدم استفاده از آموزش مشترک چندزبانه اشاره کرد. همچنین، تعمیم مدل‌ها به داده‌های دنیای واقعی با توزیع‌های متفاوت همچنان یک چالش باز باقی می‌ماند.

---

۹ . نتیجه‌گیری و کارهای آینده

در این پروژه، یک چارچوب چندزبانه مبتنی بر Transformer برای تشخیص کلاهبرداری مخابراتی ارائه شد که بر پایهٔ دیتاست CHIFRAUD و دو دیتاست تکمیلی انگلیسی و عربی بنا شده است. نتایج نشان داد که این رویکرد قادر به دستیابی به عملکرد بالا حتی در شرایط عدم توازن کلاسی است.

در آینده، می‌توان با استفاده از مدل‌های بزرگ‌تر، آموزش چندزبانهٔ مشترک و بهبود تنوع داده‌ها، عملکرد سیستم را بیش از پیش ارتقا داد. همچنین، بررسی ملاحظات عملیاتی برای استقرار این مدل‌ها در سامانه‌های واقعی می‌تواند گام بعدی این خط پژوهشی باشد.