Noise Detection fehlerhafter Dokumenten-Labels in Dokumenten der Rückversicherung mittels Machine Learning

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Fabian Creutz

Referent: Prof. Dr. Markus Döhring | Korreferentin: Prof. Dr. Antje Jahn
Darmstadt University of Applied Sciences

Einleitung

Die manuelle Qualitätssicherung von Dokumenten in Unternehmen wird durch steigende Datenaufkommen zunehmend komplex, zeitaufwendig und fehleranfällig. Falsch eingeordnete Dokumente mit Labelrauschen (Label Noise) führen in der Rückversicherung zu Such- und Korrekturaufwand. Maschinelles Lernen, speziell das Label-Noise-Filtering, bietet Möglichkeiten, solche Fehlkategorisierungen automatisiert zu identifizieren.

Frühere Forschungsansätze konzentrierten sich hierzu auf Ensemble-basierte Techniken wie z.B. den High Agreement Random Forest Filter [1] sowie das in [2] vorgestellte NoiseRank-Verfahren, wodurch für jede Instanz ein Noise-Rating entsteht. Aktuelle Entwicklungen zeigen eine breitere Methodenvielfalt: In [3] wird der kj-NN Algorithmus vorgestellt, welcher semantisch unstimmige Dokumente innerhalb einer Dokumentengruppe erkennt. Sequenzielle Ensemble-Filter wie in [4] bieten einen iterativen Ansatz, welcher die Feature-Selektion während des Noise-Detektionsprozesses als wichtig hervorhebt.

Diese Arbeit untersucht verschiedene Methoden zur Identifikation von Label-Noise und analysiert dokumentspezifische Merkmale zur Erkennung von Fehlkategorisierungen. Daraus ergeben sich folgende zentrale Fragestellungen:

- Wie effektiv sind verschiedene Machine-Learning-Methoden bei der Identifizierung fehlkategorisierter PDF-Dokumente im Rückversicherungskontext? Welche Unterschiede zeigen sich in der Präzision der untersuchten Methoden?
- Welche Features können aus den Daten generiert werden, um eine möglichst hohe Genauigkeit und Zuverlässigkeit bei der Erkennung falsch kategorisierter Dokumente zu erreichen? Welche Features eignen sich gut, welche schlecht?

Angewandte Label-Noise-Filtering Methoden

Class Noise Detection and Classification (CNDC)
Ein von Nematzadeh et al. entwickeltes, mehrstufiges Konzept, um Label-Noise zu filtern. Der erste Schritt besteht aus einem Ensemble verschiedener klassischer Klassifikationsalgorithmen, welche auf den Trainingsdaten trainiert wurden. Anschließend werden die Modelle auf künstlich verrauschte Testdaten angewandt. Durch ein Majority-Voting der Klassifikatoren erfolgt eine erste Unterteilung in schwaches oder starkes Rauschen. Der zweite Schritt umfasst eine Distanzfilterung zu rauschfreien Daten anhand der durchschnittlichen euklidischen Distanz im Rahmen einer Nachbarschaftsbetrachtung [5].

Selection Enhanced Noisy Label Training (SENT)
Ein von Wang et al. entwickeltes, mehrstufiges Framework, um iterativ verrauschte Instanzen auszusortieren. Die zentralen Aspekte dieses Frameworks nutzen zwei Hauptschritte: Erst wird die Rauschverteilung aus Trainingsdaten mithilfe des Sprachmodells BERT [6] gelernt und auf Testdaten übertragen (Noise Transfer). Anschließend trainiert ein zweiter Klassifikator mit Modellsignalen, welche aus den Testdaten berechnet werden. Abschließend erfolgt ein Transfer des zweiten Modells, um iterativ verrauschte Instanzen im Trainingsdatensatz zu identifizieren und zu entfernen (Selection Learning) [7]. Aufgrund von Hardwareeinschränkungen wurde für diese Arbeit das BERT-Modell durch ein zweischichtiges neuronales Netz ersetzt.

Sortierstrategien basierend auf Unsicherheitsmetriken
Ein öffentlich verfügbares BERT-base-uncased-Modell wurde gefinetunt und auf die Daten angewandt, um den Dokumententypen zu klassifizieren. Auf Basis der Klassenwahrscheinlichkeiten des Modells wurden verschiedene Unsicherheitsmetriken berechnet und anhand dieser eine Sortierung der Instanzen durchgeführt. Berechnet wurden die Entropie, die First-Second-Distance [8] sowie die Lowest Ground Truth [8]. Anhand der Sortierung erfolgte jeweils die Betrachtung der ersten 5%, 7% und 10% der Instanzen.

Datensatz

Als Datengrundlage dienten PDF-Dateien aus dem Dokumentenmanagementsystem der R+V Re, welche mittels Python extrahiert wurden. Hierbei wurden 49.637 Dokumente extrahiert, welche insgesamt 51 unterschiedliche Labels aufweisen und zu 13 allgemeinen Labels aggregiert wurden.

Entwickelte Features

Im Rahmen dieser Arbeit wurden unterschiedlichste Features entwickelt, welche in sinnvolle Gruppierungen unterteilt wurden: (1) strukturelle Merkmale wie z.B. die Höhe, Breite und Größe eines Dokuments, (2) textbasierte Merkmale wie z.B. Term Frequency - Inverse Document Frequency Scores, Textstatistiken oder die Verteilung der Textblöcke im Dokument und (3) weitere Features wie z.B. die Topicwahrscheinlichkeiten eines Dokuments oder Embeddings. Insgesamt ergeben sich nach der Datenvorverarbeitung 912 Eingabefeatures, welche zum Training der Basis-Modelle des CNDC sowie des SENT-Frameworks verwendet werden. Das Sprachmodell BERT, welches in der Sortierstrategie angewandt wird, erhält für jedes Dokument die ersten 512 Tokens des extrahierten Rohtexts als Eingabe. Zur Evaluierung der Feature Importance wurden für die erstellten Feautres der Mean Decrease in Impurity sowie der Mean Decrease in Accuracy berechnet. Weiterhin wurden vier Feature-Gruppierungen erstellt, in welchen jeweils die wichtigsten 250 und 500 Features anhand jeder dieser beiden Metriken enthalten sind.

Ergebnisse

Class Noise Detection and Classification (CNDC) Die durchschnittliche Präzision aus fünf Durchläufen der drei besten CNDC-Modelle auf unterschiedlichen Rauschniveaus künstlich verrauschter Testdaten sind in der unten gezeigten Tabelle aufgeführt. Diese wurden jeweils mit den gruppierten textbasierten sowie den 250 und 500 wichtigsten Features anhand der Mean Decrease in Impurity (MDI) trainiert. Zudem wurde das Experiment für unterschiedliche k auf den Testdaten wiederholt. Das Ergebnis zeigt, dass für die Modelle mit textbasierten Features niedrige Werte für k zu besseren, für die beiden anderen Modelle größere Werte für k zu besseren Ergebnissen führen. Die alleinige Verwendung der strukturierten Merkmale sowie der Embedding-Features führte zu deutlich schlechteren Präzisionswerten über alle Rauschniveaus hinweg.

Feature Loadout	künstliches Rauschniveau (in %)				
	10	20	30	40	50
textbasierte Features	0.548	0.736	0.832	0.875	0.913
wichtigsten 500 Features anhand der MDI	0.503	0.697	0.792	0.840	0.898
wichtigsten 250 Features anhand der MDI	0.489	0.674	0.781	0.840	0.897

Table 1. Präzision der 3 besten CNDC-Modelle mit k=10 für die Distanzbetrachtung.

Eine Übertragung des besten Modells auf die verrauschten Trainingsdaten resultiert in einer Präzision der Rausch-Vorschläge von 73% mit 1551 korrekt erkannten Rauschinstanzen. Aus den Experimenten geht hervor, dass die drei besten CNDC-Modelle ab einem Rauschniveau von 20% in der Lage sind, Rauschen auf den Testdaten mit einer akzeptablen Präzision zu erkennen. Aus den Ergebnissen geht hervor, dass das Modell, welches mit textbasierten Features trainiert wurde, die besten Ergebnisse liefert.

Selection Enhanced Noisy Label Training (SENT) Die SENT-Frameworks basieren auf einem verrauschten Trainingsdatensatz, welcher iterativ gesäubert wird. Hierzu wurden je Framework fünf SENT-Epochen durchgeführt, deren Ergebnisse aus Epoche fünf in der nachfolgenden Tabelle aufgeführt sind. Die 3 besten SENT-Frameworks sind in der Lage, ähnlich viele verrauschte Instanzen auf den Trainingsdaten zu identifizieren wie die CNDC-Modelle, allerdings mit deutlich schlechterer Präzision.

Feature Load	True Positive	False Positive	Präzision
wichtigsten 250 Features anhand der MDI	1487	1570	0.49
alle entwickelten Features	1488	1590	0.48
wichtigsten 500 Features anhand der MDI	1397	1567	0.47

Table 2. Präzision der 3 besten SENT-Frameworks auf Trainingsdaten nach 5 SENT-Epochen.

In der Evaluation ist aufgefallen, dass die Frameworks hierbei in den ersten beiden SENT-Epochen ihre jeweils besten Präzisionswerte erreichen, welche in den nachfolgenden Epochen stark einbrechen. Die Präzisionswerte der drei besten Frameworks unterscheiden sich kaum und befinden sich unter 50%. Zusammenfassend kann man erkennen, dass sich die aus den Testdaten abgeleiteten Signale des zweischichtigen neuronalen Netzwerks nicht gut als Features für eine Label-Noise-Klassifikation eignen. Es bleibt festzuhalten, dass die modellbasierten Features für diesen Anwendungsfall besser als Noise-Indikation geeignet sind und sich durchaus dazu eignen, die Gesamtdatenmenge auf eine Fokusmenge einzugrenzen, welche man im Nachgang genauer betrachten kann.

Sortierstrategien basierend auf Unsicherheitsmetriken

Zunächst konnte die Resistenz von BERT [6] gegenüber Label-Rauschen durch diese Arbeit bestätigt werden. Das auf verrauschten Daten trainierte Sprachmodell erzielt auf rauschfreien Testdaten eine Klassifikationsgenauigkeit von 0.85 und einen F1-Score von 0.79. Im Rahmen der Sortierstrategien, um Label-Noise zu identifizieren hat sich gezeigt, dass sich eine Sortierung anhand der Lowest Ground Truth am besten eignet, um verrauschte Instanzen zu finden. Hierbei weist ein Cutoff-Wert von 5% die höchste Präzision von ca. 50% auf (mit 1191 True Positives). In den durchgeführten Experimenten konnte für alle Strategien beobachtet werden, dass eine Erhöhung des Cutoffs auf 7% oder 10% mit einem großen Anstieg gefundener True Positives einhergeht, allerdings auf Kosten der Präzision. Somit identifizierte die LowGT-Strategie mit einem Cutoff von 10% insgesamt 1841 verrauschte Instanzen in den Trainingsdaten, jedoch mit einer Präzision von lediglich 38.5%. Die beiden anderen Strategien erwiesen sich über alle Cutoff-Werte hinweg als schlechter. Es bleibt festzuhalten, dass sich diese Strategie zwar verrauschte Instanzen identifizieren kann, allerdings nicht verlässlich und mit einer niedrigen Präzision. Ähnlich wie bei den SENT-Frameworks eignen sich die unterschiedlichen Unsicherheitsmetriken viel mehr als Indikation, um die Gesamtmenge der Daten auf eine Fokusmenge einzuschränken, welche potenziell verrauschte Instanzen enthält.

Referenzen

[1] B. Sluban, D. Gamberger, and N. Lavra, "Advances in class noise detection," in *European Conference on Artificial Intelligence*, 2010.

[2] B. Sluban, D. Gamberger, and N. Lavrač, "Ensemble-based noise detection: noise ranking and visual performance evaluation," in *Data Mining and Knowledge Discovery*, 2014.

[3] E. Fouché, Y. Meng, F. Guo, H. Zhuang, K. Böhm, and J. Han, "Mining text outliers in document directories," in *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 152–161, 2020.

[4] D. Guan, K. Chen, G. Han, S. Huang, W. Yuan, M. Guizani, and L. Shu, "A novel class noise detection method for high-dimensional data in industrial informatics," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 2181–2190, 2021.

[5] Z. Nematzadeh, R. Ibrahim, and A. Selamat, "Improving class noise detection and classification performance: A new two-filter cndc model," *Applied Soft Computing*, vol. 94, p. 106428, 2020.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[7] Z. Wang, Z. Lin, J. Wen, X. Chen, P. Liu, G. Zheng, Y. Chen, and Z. Yang, "Learning to detect noisy labels using model-based features," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, (Abu Dhabi, United Arab Emirates), pp. 5796–5808, Association for Computational Linguistics, Dec. 2022.

[8] D. Henter, A. Stahl, M. Ebbecke, and M. Gillmann, "Classifier self-assessment: active learning and active noise correction for document classification," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 276–280, 2015.

[9] S. Zhong, W. Tang, and T. Khoshgoftaar, "Boosted noise filters for identifying mislabeled data," 01 2005.