supper&supper

NLP – Verstehen wissenschaftlicher Sprache

Projektziel

Es gibt zwei verschiedene Arten von Genmutationen - eine, welche zum Wachstum von Krebstumoren beiträgt und eine zweite, die zu neutralen Genmutationen beiträgt. Derzeit wird die Klassifizierung solcher Genmutationen manuell durch klinische Pathologen durchgeführt. Die Klassifizierung erfolgt auf Grundlage von Erkenntnissen aus wissenschaftlicher klinischer Literatur, welche unter hohem Zeitaufwand gesichtet wird.

Der Use Case beschreibt die Entwicklung eines Algorithmus zur automatischen Unterscheidung dieser beiden verschiedenen Arten von Mutationen mittels Natural Language Processing (NLP) zur Analyse wissenschaftlicher Artikel. Dieser Ansatz kann klinischen Pathologen viel Zeit und Energie ersparen. Durch die Kenntnis der Mutationsklasse können Ärzte Krankheiten genauer diagnostizieren und Patienten gezielter behandeln.

Zur Verfügung gestellte Daten

Es standen ein Trainingsdatensatz mit 3321 Samples und ein Testdatensatz mit 368 Samples zur Verfügung. Die Trainingsdaten enthalten wissenschaftliche Volltextpublikationen samt zugehörigem Gen und einer von Onkologen ausgewerteten Mutationsklasse. Die Gene sind dabei in insgesamt 9 verschiedene Klassen eingeteilt. Die Testdaten haben die gleiche Struktur wie die Trainingsdaten, jedoch ohne die Mutationsklasse. Die Klassifizierung der Testdaten wird durch das zu entwickelnde Model vorhergesagt und darauf basierend die Leistung des Modells bewertet.

Herausforderungen

Die Klassen der Trainingsdaten sind sehr unausgeglichen, wobei die Klassen 3, 8 und 9 nur 5% und die Klassen 7 und 4 über 50% der Gesamtdaten ausmachen. Die Anzahl der Trainingsproben ist für eine 9-klassige NLP-Klassifizierung insgesamt relativ gering. Die Verwendung vortrainierter Modelle könnte aufgrund der hohen Spezifität der wissenschaftlichen Literatur eingeschränkt sein. Die Texte selbst haben sehr variablen Umfang und reichen von nur 80 bis über 70.000 Wörter. Eine Vereinheitlichung der Texte kann zum Verlust wichtiger Informationen führen.

Angewandte Methoden

Bevor die Texte in ein Modell eingegeben werden, sind verschiedene Vorverarbeitungsschritte erforderlich. Die Texte werden in eine Liste von sogenannten "Token", d.h. Einzelwörtern, umgewandelt. Unerwünschte Token, wie Stoppwörter oder Wörter, welche nur einmal vorkommen, werden herausgefiltert. Anschließend wird ein "Word Embedding" durchgeführt - die Umwandlung der Token in Zahlen. Drei verschiedene Methoden wurden für die Klassifizierung verwendet: *Bag of Words, TF-IDF* und *Word2Vec.* Verglichen wurden zusätzlich zwei Word2Vec-Modelle: ein vortrainiertes Modell von Google und ein von Grund auf selbsttrainiertes Modell.

Für die folgende Klassifizierung wurden drei klassische Maschinenlernalgorithmen verwendet und weiter verglichen: logistische Regression, zufällige Waldklassifizierung sowie Support-Vektor-Klassifizierung.

Zusätzlich wurde BERT mit einem Fine-Tuning Ansatz für das Word-Embedding und die Modellierung eingesetzt. BERT ist ein umfassendes NLP Framework von Google und hat viele bisherige Rekorde gebrochen, was es zum aktuellen State-of-the-Art Ansatz in der Verarbeitung natürlicher Sprache macht. Es kann für eine Vielzahl von Aufgaben eingesetzt werden und liefert konstant gute Ergebnisse.

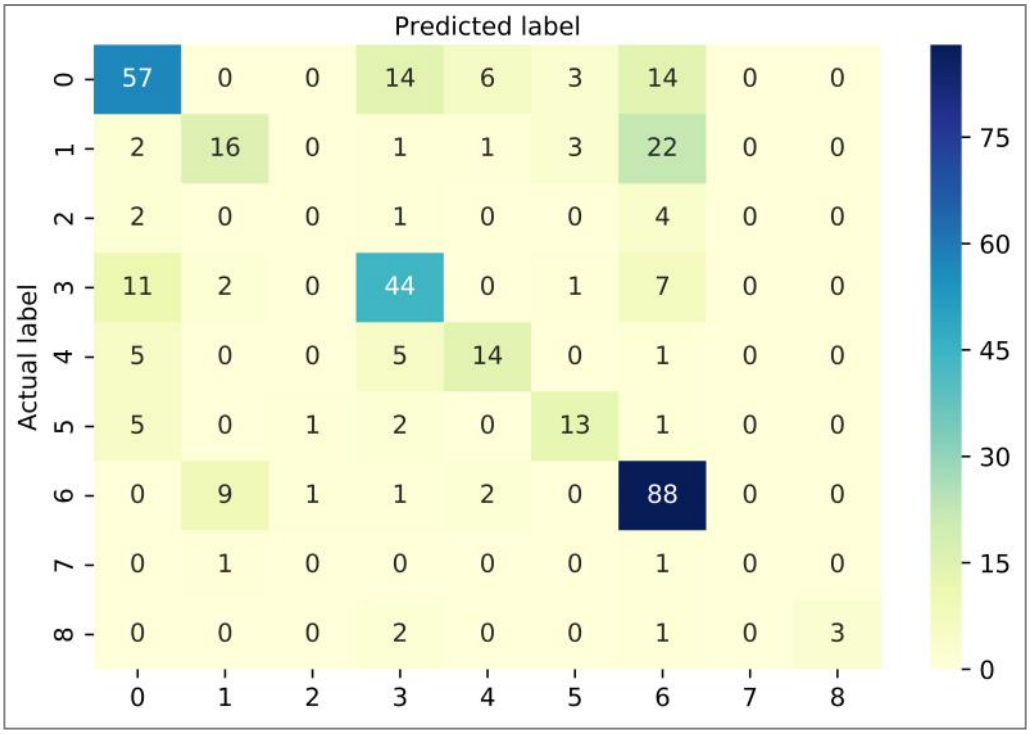
Projektergebnis

Für die Auswertung wurde der Multi-Class Log Loss (MCLL) der Vorhersagewahrscheinlichkeit verwendet. Je niedriger die MCLL, desto besser fällt die Leistung eines Modells aus. Die Baseline wurde mit zufälligen Wahrscheinlichkeiten der Klasse erstellt. Von den drei verwendeten Modellen war die Random Forest Klassifizierung die beste unter allen Methoden. Von den verwendeten Word-Einbettungstechniken liefert das selbst trainierte *Word2Vec* kontinuierlich die besten Ergebnisse. Die Kombination aus selbst trainiertem W2V und Random Forest Modell bietet die beste Leistung. Mit dieser Kombination wurde eine Genauigkeit von 63,5% erreicht (MCLL 0,94).

Überraschenderweise konnten vortrainierte Modelle von BERT die Ergebnisse der oben genannten Methoden nicht verbessern. Mögliche Erklärungen sind der Bedarf an mehr Trainingsdaten und eine allgemeine schlechte Leistung von vortrainierten Modellen zu sehr spezifischen Texten, wie der verwendeten wissenschaftlichen Literatur. Verbesserungspotenzial liegt in der Generierung von mehr Trainingsdaten und Hyperparameter-Optimierung. Aufgrund der immensen Laufzeit wurde für *Bag of Words* keine Support-Vector-Klassifizierung durchgeführt.

	Bag of Words	TF-IDF	Vor-trainiertes W2V	Selbst-trainiertes W2V	BERT	Baseline
Logistic Regression	1.08	1.03	1.40	0.99	1.13	2.54
Random Forest Classifier	1.02	1.01	0.98	**0.94**		
Support Vector Classifier	—	1.01	1.22	0.97		

Multi-Class Log Loss der getesteten Modelle und die Einbettungsmethoden (niedriger = besser)



Confusion Matrix der Klassifizierung mittels Random Forest und Daten des selbsttrainierten *Word2Vec* Modells

Das Ergebnis dieses Projekts sind Vorhersagen über genetische Mutationsklassen, die auf wissenschaftlichen Artikeln basieren. Wenn eine Genmutation mit zugehöriger Variation für einen Patienten betrachtet wird, kann mit Hilfe zugehöriger wissenschaftlicher Artikel die potentielle Mutationsklasse für diese Genmutation vorhergesagt oder zumindest eingegrenzt werden. Der zuständige Onkologe oder Pathologe kann diese Informationen auch für eine Diagnose nutzen. Auf diese Weise kann die Heilungswahrscheinlichkeit erhöht werden, indem eine rechtzeitige Behandlung durch die frühzeitige Erkennung von Krankheiten ermöglicht wird.