Kurzbeschreibung, Idee, Gegenstand und Zielsetzung der Thesis

Thema : Textanalyse mit Convolutional Neural Networks

Student: Michael Federer

Betreuende Dozentin: Ursula Deriu

Datum: 12. Juni 2018

Vorschlag 1

Kurzbeschreibung der Idee

[2] Bei der Analyse von Kurztexten besteht in der Regel das Problem, dass die enthaltenen Informationen / Worte sehr spärlich vorliegen und oft mehrdeutig interpretiert werden können. Im Extremfall sind keinerlei Worte enthalten, welche zur Unterstützung des Trainings eines Modells angewendet werden können. Bei den bereits länger existierenden Vektorraummodellen wurden bereits einige Anstrengungen unternommen, um verbesserte Verfahren / Methoden zur Klassifizierung von Kurztexten zu entwickeln. Da für die Textklassifizierung mittels CNN ein Textkörper zuerst in seine Vektorraumrepräsentation überführt werden muss, wäre es interessant einen [2] neuen Ansatz, der die nachträgliche Bildung von semantischen Cliquen vorschlägt, zu testen.

Konkret soll dieses Verfahren an der Klassifizierung von MEDLINE/PubMed Abstracts getestet werden und die Güte der Resultate mit [1] Resultaten von CNN+Word2Vec Experimenten verglichen werden. Es wird erwartet, dass die Klassifizierungsperformance mit dem vorgeschlagenen Verfahren zunimmt. Die [5] MEDLINE/PubMed-Metadatenbank wird für die Bestimmung der Klassifizierungsperformance eines Verfahrens oft herangezogen, da die dort publizierten Arbeiten für die Medizin von grossem wissenschaftlichem Interesse sind und eine exakte Zuordnung dieser, den Umgang massiv beschleunigt. Zudem sind die bereits existierenden Datensätze mit einem [3, 4] Medical Subject Heading versehen, welche das Training von Vektorräumen ermöglichen würde.

Gegenstand der Thesis

- Vorprozessierung (Vektorraumrepräsentation) von MEDLINE/PubMed Abstracts nach dem vorgeschlagenen Verfahren von [2]
- Klassifizierung von PubMed-Publikationen anhand ihres Abstracts mittels Semantic-CNN
- Vergleich der Klassifizierungsperformance von Semantic-CNN vs. CNN+Word2Vec

Zielsetzung der Thesis

Untersuchung von Semantic-CNN zur Eignung der Dokumentenklassifikation von MEDLINE/PubMed-Dokumenten anhand ihres Abstracts.

Vorläufige Freigabe / Ablehnung der konkreten Idee / Themenstellung

???

Quellenverzeichnis

[1] HUGHES Mark et al. : Medical Text Classification using Convolutional Neural Networks

[2] WANG Peng et al. : Semantic Clustering and Convolutional Neural Network for Short Text Categorization

[3] MeSH (Medical Subject Heading) : https://de.wikipedia.org/wiki/Medical_Subject_Headings

[4] MEDLINE/PubMed – MeSH Terms: https://www.nlm.nih.gov/bsd/mms/medlineelements.html#ab

[5] MEDLINE/PubMed: https://www.ncbi.nlm.nih.gov/pubmedhealth/PMHT0029900/

Vorschlag 2

Kurzbeschreibung der Idee

[2] Bei der Analyse von Kurztexten besteht in der Regel das Problem, dass die enthaltenen Informationen / Worte sehr spärlich vorliegen und oft mehrdeutig interpretiert werden können. Im Extremfall sind keinerlei Worte enthalten, welche zur Unterstützung des Trainings eines Modells angewendet werden können. Bei den bereits länger existierenden Vektorraummodellen wurden bereits einige Anstrengungen unternommen, um verbesserte Verfahren / Methoden zur Klassifizierung von Kurztexten zu entwickeln. Da für die Textklassifizierung mittels CNN ein Textkörper zuerst in seine Vektorraumrepräsentation überführt werden muss, wäre es interessant einen [2] neuen Ansatz, der die nachträgliche Bildung von semantischen Cliquen vorschlägt, zu testen.

Konkret soll dieses Verfahren an der Klassifizierung von Kurztexten / -nachrichten getestet werden und die Güte der Resultate mit den Resultaten von [1] verglichen werden. Die Kurztexte sind auf ihren Gehalt und Art an vergiftenden Bestandteilen zu prüfen und entsprechend zu klassifizieren. Ein solcher Klassifizierer könnte beispielsweise in den sozialen Medien (Facebook etc.) genutzt werden, um die Benutzer vor sexistischen, rassistischen oder in sonst irgendeiner Weise persönlichkeitsverletzenden Kommentare zu schützen. Das öffentliche Interesse an effizienten Instrumenten zur Bekämpfung solcher Kommentarbeiträge scheint gross zu sein.

Gegenstand der Thesis

- Vorprozessierung (Vektorraumrepräsentation) von Kurznachrichten nach dem vorgeschlagenen Verfahren von [2]
- Klassifizierung von [3] Kurznachrichten mittels Semantic-CNN
- Vergleich der Klassifizierungsperformance von Semantic-CNN vs. CNN_{fix}-Word2vec

Zielsetzung der Thesis

Untersuchung von Semantic-CNN zur Eignung für die Toxic Comment Classification auf Basis des gelabelten [3,4,5] Wikipedia Talk Corpus den [1] verwendet hat.

Vorläufige Freigabe / Ablehnung der konkreten Idee / Themenstellung

???

Quellenverzeichnis

[1] GEORGAKOPOULOS Spiros et al. : Convolutional Neural Networks for Toxic Comment Classification

[2] WANG Peng et al. : Semantic Clustering and Convolutional Neural Network for Short Text Categorization

[3] Kaggle – Toxic Comment Classification Challenge : https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[4] Wikipedia Detox: https://meta.wikimedia.org/wiki/Research:Detox/Data_Release

[5] Wikipedia Talk Corpus: https://figshare.com/projects/Wikipedia_Talk/16731