15-115-R "Text Mining als Werkzeug zur Überwindung der Wissensfragmentierung in der psychiatrischen Forschung (PsyMine)" CHF 188'892.-

Prof. Meichun Mohler-Kuo; Drs. Simon Foster, Fabio Rinaldi, UZH

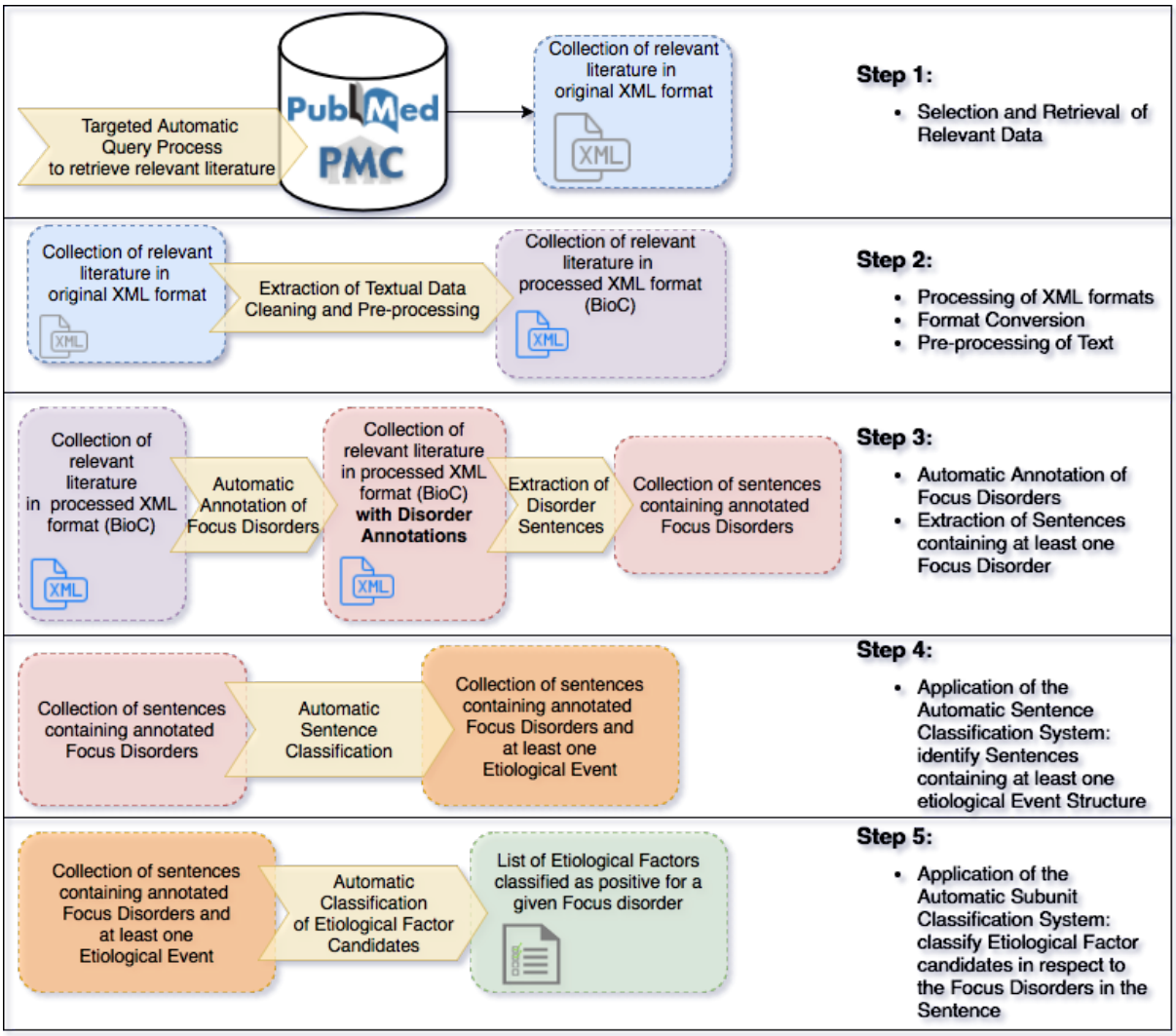
Psychische Erkrankungen sind für einen substantiellen Anteil der weltweiten Krankheitslast verantwortlich. Aufgrund der Menge an wissenschaftlicher Literatur und deren Verteilung über diverse verschiedene Disziplinen hinweg (z.B. Neurobiologie, Psychiatrie, und Soziologie), ist es für Forschende und klinisch tätige Personen jedoch schwierig, sich einen Überblick über das Ursachengefüge der Erkrankungen zu verschaffen. Das Ziel des "PsyMine-Projekts" bestand in der Entwicklung von Text-Mining-Applikationen, welche die automatische Erkennung von jenen Textstrukturen in der wissenschaftlichen Literatur erlauben, die einen Zusammenhang zwischen einem Risikofaktor und einer psychischen Erkrankung zum Ausdruck bringen. Dadurch wird es möglich, grosse Mengen an Fachliteratur automatisiert zu verarbeiten und relevante Risikofaktoren über wissenschaftliche Disziplinen hinweg zu identifizieren. Im Fokus des Projekts standen dabei Internalisierungsstörungen, insbesondere Depressions- und Angststörungen, da diese zu den häufigsten psychischen Erkrankungen gehören.

Als Quelle relevanter wissenschaftlicher Literatur wurden "PubMed" und "PubMed Central" verwendet. "PubMed" ist die weltweit am häufigsten verwendete Datenbank und Suchmaschine für medizinische Literatur mit zurzeit rund 26 Millionen Literatureinträgen. "PubMed Central" ist eine mit "PubMed" verbundene Datenbank mit frei zugänglichen Volltexten. Beide Datenbanken werden vom "National Center for Biotechnology Information" (NCBI) an der "US National Library of Medicine" (NLM) betrieben und zur Verfügung gestellt.

Der Kern des "PsyMine-Projekts" bestand in der Entwicklung einer vollständig automatisierten Pipeline, welche relevante Literatureinträge in "PubMed" identifiziert, die zugehörigen Abstracts und – soweit zugänglich – Volltexte herunterlädt, diese in ein günstiges Format konvertiert ("BioC"-Format), anschliessend Nennungen der relevanten psychischen Erkrankungen in den formatierten Abstracts und Volltexten annotiert, alle Sätze mit entsprechenden Annotationen extrahiert, die Sätze klassifiziert in "Zusammenhangsaussage" versus "keine Zusammenhangsaussage", und schliesslich die Risikofaktoren aus den als "Zusammenhangsaussage" klassifizierten Sätzen extrahiert. Die Pipeline ist schematisch in Abbildung 1 "PsyMine Pipeline" dargestellt.

Im Rahmen des Projekts wurden diverse Teilprobleme bearbeitet, um die einzelnen Schritte der Pipeline zu implementieren. Dazu gehörten die Erstellung eines systematischen Vokabulars für psychische Erkrankungen und Risikofaktoren; die Entwicklung eines Korpus von 175 Abstracts, in denen psychische Erkrankungen, Risikofaktoren und linguistische Strukturen, welche Zusammenhangsaussagen anzeigen, manuell annotiert wurden; die Entwicklung von Programmen für die Annotation, Formatierung, und Evaluation des Korpus; die Sicherstellung der Konsistenz der Annotationen im Textkorpus; die linguistische und quantitative Analyse des Korpus; die Entwicklung von Programmen für die Identifizierung und die Aufbereitung relevanter Abstracts und Volltexten; die Entwicklung und Evaluation statistischer Modelle zur Klassifizierung von Sätzen in "Zusammenhangsaussage" versus "keine Zusammenhangsaussage"; sowie die Entwicklung und Evaluation statistischer Modelle zur Identifizierung der Risikofaktoren innerhalb der Zusammenhangsaussagen.

Das Projekt sowie verschiedene Zwischenresultate wurden an mehreren internationalen Konferenzen und eingeladenen Vorträgen vorgestellt, u.a. am "Biomedical Linked Annotation Hackathon" (16.-20. November 2015, Ito, Japan), an der "10th International Conference on Language Resources and Evaluation (LREC) " (23.-28. Mai 2016, Portoroz, Slowenien), am "13th Scientific Meeting of the Swiss Society of Psychiatric Epidemiology" (24. Juni 2016, Zürich), und an der "28th Annual Convention of the Association for Psychological Science" (26.-29. Mai 2016, Chicago, USA).
Zum Zeitpunkt des Projektendes wurden 1'111 Wortstrukturen, die sich auf psychische Erkrankungen beziehen, sowie 3793 Wortstrukturen, die sich auf Risikofaktoren beziehen, ins Vokabular aufgenommen. Insgesamt 217'737 relevante Literatureinträge wurden in "PubMed" identifiziert, für welche 165'925 Abstracts und 9'317 Volltexte heruntergeladen und verarbeitet wurden. Aus diesem Textmaterial wurden 55'787 potentielle Zusammenhangsaussagen extrahiert.



Das "PsyMine-Projekt" hat gezeigt, dass Risikofaktoren von psychischen Erkrankungen automatisiert aus der wissenschaftlichen Literatur extrahiert werden können. Der Pipeline-Prototyp bestätigte damit das Potential von Text-Mining-Applikationen für die Erforschung von psychischen Erkrankungen. Gleichzeitig kann der Prototyp noch in vielerlei Hinsicht verbessert werden, insbesondere hinsichtlich des Web-Interfaces, welches den einfachen Nutzerzugang zu den extrahierten Informationen ermöglicht und erst vorläufig installiert wurde. Die Projektmitglieder werden weiter zusammenarbeiten, um die Pipeline zur breiten Anwendung in der Erforschung der psychischen Erkrankungen zu bringen.