Knowledge discovery from patient forums: gaining novel medical insights from patient experiences

Dirkson, A.R.

SUMMARY

Patients share valuable advice and experiences with their peers in online patient discussion groups. These uncensored experiences can provide a complementary perspective to that of the health professional and thereby yield novel hypotheses which could be tested in further rigorous medical research. This thesis focuses on the development of automatic extraction methods to harvest these patient experiences from online patient forums using text mining techniques. We also examine the complementary value of these patient-reported outcomes to traditional sources of medical knowledge for scientific hypothesis generation. Specifically, we focus on the extraction of adverse drug events (i.e. side effects) and coping strategies for dealing with adverse drug events.

In the first part, we investigated how spelling mistakes in medical social media messages can be reduced. We studied which unsupervised spelling correction method is most suitable for correcting spelling mistakes in medical social media data without losing valuable information due to false positives (domain-specific terms or layman's terms that are corrected because they are not in the dictionary). We also examined how posts containing patient experiences can best be distinguished from those that do not to weed out irrelevant posts. This helped to select the discussion threads that are most likely to contain adverse drug events (ADEs) and provided insight into the different types of patient experiences shared on the forum. In a third study, we showed that despite the fact that relevant posts cluster together, incorporating the structure of the conversation into state-of-the-art text classification models did not help to identify relevant posts.

In the second part, we addressed challenges presented by the extraction of text snippets containing patient-reported ADEs. We tested the efficacy of default Transformer models, including the popular BERT model, for this task and evaluated the vulnerability of BERT models to being fooled by variation in the input data. We also tackled the challenge of discontinuous entities, which can be either composite (e.g. "*hand* and *foot pain*") or disjoint (e.g. "*eyes* are feeling *dry*"). We presented a more flat, continuous representation of these entities that can benefit end-to-end extraction of ADEs.

In the third part, we showcased a novel task: the extraction of coping strategies for adverse drug responses. We presented the first ontology for coping strategies, compared the success of different conceptualizations of this task, and showed that automatically derived coping strategies from an online patient forum could be used for hypothesis generation.

In the fourth part, we described a case study on a specific patient forum for Gastro-Intestinal Stromal Tumor (GIST) patients and demonstrated the value of extracting ADE from patient forum posts for post-market drug monitoring. We showed that adverse drug events can be extracted from patient forum messages with sufficient success to enable the discovery of novel ADEs, long-term ADEs, and an indication of which ADEs are most important to patients. A comparison of these results with ADEs reported by GIST patients in a survey revealed that automatically extracted ADEs from patient forum data

can be used to select the most appropriate questionnaire for the patient population and to keep questionnaires up to date. To better understand the limitations of knowledge discovery from patient forum data, we also investigated how representative the online patient population of GIST patients is. We found that patients in relatively better condition are generally under-represented on the patient forum.

Our work offers a starting point for knowledge discovery from online patient forums and its use as a complementary data source for hypothesis generation. Future work will need to elucidate to what extent the complementary value of patient knowledge may differ between different types of disorders, such as between rare and more common disorders and between chronic and more acute disorders.