

Automatically Analyzing Online Patient Experience Data with Natural Language Processing: An Instrument to Investigate Health Status and Help-Seeking Factors in Patients with Obesity

Matthias Hartung¹, Nathalie Schwering¹, Yannick Loonus¹, Philipp Cimiano^{1,3}, Anna A. Jäger², Ben Collins²

¹Semalytix GmbH, Bielefeld, Germany

²Boehringer Ingelheim International GmbH, Ingelheim, Germany

³Semantic Computing Group, Faculty of Technology, Bielefeld University, Germany

AIMS: Self-reported patient narratives from social media provide a rich source of information towards understanding the real-world experience of patients from an unsolicited first-hand perspective. Due to volume and heterogeneity, automated methods are required for transforming patient-generated content into real-world data to support patient-focused drug development. Using obesity as an example, this study explores the benefit of such data sources for gathering insights into the real-world experience of a disease as reported by large online patient populations.

METHODS: Purposive sampling yielded 1.8M unique posts from obesity-related online conversations in six social media sources between January 2018 and February 2020. Using state-of-the-art technology from natural language processing, this corpus was reduced to approx. 22,000 narratives by patients with obesity, a subset of which (n=6,124) contains explicit references to symptoms and co-morbidities. Within this sub-corpus, 1,424 documents were automatically mapped to speakers from four BMI classes (25-29.9; 30-34.9; 35-39.9; >40); 3,400 documents were automatically assigned to thematic clusters focusing on doctor-patient relationships (n=2,863) and moments of motivation for trying to lose weight (n=830).

RESULTS: Quantitative analysis reveals marked differences between the distributions of symptoms and co-morbidities across BMI classes. Cardio-metabolic conditions (diabetes mellitus, hypertension) are consistently over-expressed in BMI classes 35-39.9 (12.8% vs. 8.8%, 9.0% vs. 3.0%, respectively) and >40 (13.8% vs. 8.8%, 9.8% vs. 3.0%, respectively) compared to the overall data set. Depression is strongly over-expressed in higher BMI classes (10.3% in BMI 35-39.9 vs. 7.2% all; 17.0% in BMI >40 vs. 7.2% all), thus being the most frequently reported burden by morbidly obese patients. Qualitative interpretation of thematic clusters shows that the majority of motives for help-seeking are of intrinsic and positive orientation. Supportive behaviors vs. blaming are found as main drivers of a positively or negatively perceived doctor-patient relationship.

CONCLUSIONS: This study corroborates the potential of automatically analyzing patient-generated online data to complement and extend traditional instruments in market and outcomes research. In the obesity case, the approach yields confirmatory results about the bi-directional relationship of obesity and depression, while unlocking an additional data source for qualitative research into motivational factors and drivers of effective patient-doctor relationships.