



Universiteit
Leiden
The Netherlands

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

Dirkson, A.R.

Citation

Dirkson, A. R. (2022, December 6). *Knowledge discovery from patient forums: gaining novel medical insights from patient experiences*. SIKS Dissertation Series. Retrieved from <https://hdl.handle.net/1887/3492655>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3492655>

Note: To cite this publication please use the final published version (if applicable).

KNOWLEDGE DISCOVERY FROM PATIENT FORUMS

GAINING NOVEL MEDICAL INSIGHTS FROM PATIENT
EXPERIENCES

KNOWLEDGE DISCOVERY FROM PATIENT FORUMS

GAINING NOVEL MEDICAL INSIGHTS FROM PATIENT
EXPERIENCES

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op dinsdag 6 december 2022
klokke 11:15 uur

door

Anne Dirkson

geboren te Gouda, Nederland
in 1993

Promotores:

Prof. dr. ir. W. Kraaij

Prof. dr. A.J. Gelderblom

Co-promotor:

Dr. S. Verberne

Promotiecommissie

Prof. dr. A. Plaat (Leiden Universiteit)

Prof. dr. M. Spruit (Leiden Universiteit)

Prof. dr. K. Verspoor (RMIT University)

Prof. dr. M. Hoogendoorn (Vrije Universiteit)

Prof. dr. L. van de Poll (Tilburg University)



Universiteit
Leiden
The Netherlands

SIDNfonds



Printed by: Ridderprint

Front & Back: Gaby de Jong-Verwegen

This thesis was the winner of the 2022 Krijn Rietveld Memorial Innovation Award

Copyright © 2022 by A.R Dirkson

This publication was supported by the SIDN fonds

SIKS Dissertation Series No. 2022-26

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

CONTENTS

1	Introduction	1
1.1	Patient forums as a source of collective experiential knowledge	4
1.2	Using text mining to harvest collective experiential knowledge	5
1.3	Potential for rare diseases	6
1.4	Previous work focused on pharmacovigilance	7
1.5	Methodological tasks and challenges	8
1.5.1	ADE extraction	9
1.5.2	ADE normalization	11
1.5.3	Determining ADE–drug relations	12
1.6	Research questions and thesis outline	13
1.6.1	Part 1: Reducing noise in medical social media text	14
1.6.2	Part 2: Extracting adverse drug events	15
1.6.3	Part 3: Extracting recommended coping strategies.	15
1.6.4	Part 4: Complementary value of disease-specific patient forums as a source of patient-reported outcomes	16
1.7	Scientific contributions	16
1.8	Guide for the reader.	19
I	Preprocessing medical social media text	21
2	Data-driven spelling correction	23
2.1	Introduction	24
2.2	Related work	25
2.2.1	Challenges in correcting spelling errors in medical social media	25
2.2.2	Lexical normalization of social media	26
2.2.3	Lexical normalization of clinical records	27
2.3	Materials and methods	28
2.3.1	Data	28
2.3.2	Methods	29
2.4	Results	35
2.4.1	Error distribution	35
2.4.2	Spelling correction	35
2.4.3	Spelling mistake detection	39
2.4.4	Impact of corpus size	41
2.4.5	Impact of the degree of noisiness of the data	42
2.4.6	Effect on OOV rate	45
2.4.7	External validation	46
2.5	Discussion	48
2.6	Conclusions and future work	50

3	Detecting personal experiences	51
3.1	Introduction	52
3.2	Related work	52
3.3	Methods	53
3.3.1	Data	53
3.3.2	Preprocessing	53
3.3.3	Supervised classification	53
3.3.4	Topic modeling of the whole data set	54
3.4	Results	55
3.4.1	Annotated data	55
3.4.2	Classifier evaluation	55
3.4.3	Influential features	55
3.4.4	Error analysis for the best performing classifier	56
3.4.5	Frequency and content of patient narratives	57
3.5	Discussion	58
3.6	Conclusion	60
4	Conversation-aware filtering of relevant messages	61
4.1	Introduction	62
4.2	Related work	63
4.3	Methods	63
4.3.1	Models	63
4.3.2	Feature analysis	64
4.3.3	Model comparison	64
4.4	Data	64
4.5	Results	66
4.5.1	Distribution of the target class in the discussion threads	66
4.5.2	Model comparison	66
4.5.3	Analysis of selected features	68
4.6	Discussion	68
II	Extracting adverse drug effects (ADEs)	71
5	Transfer learning for ADE extraction from Twitter	73
5.1	Introduction	74
5.2	Task descriptions	74
5.3	Our approach	74
5.3.1	Preprocessing	74
5.3.2	Additional Data	74
5.3.3	Text Classification (S1 and S4)	75
5.3.4	Named Entity Recognition (S2)	75
5.3.5	Concept normalization (S3)	76
5.4	Results	76
5.5	Conclusions	78

6	Vulnerabilities of BERT for Named Entity Recognition	79
6.1	Introduction	80
6.2	Related work	81
6.3	Methods	82
6.3.1	Aim of the attacks	82
6.4	Experiments	84
6.4.1	Data	84
6.4.2	Target models	85
6.4.3	Evaluation of adversarial attacks	86
6.5	Results	86
6.5.1	Entity Context Attack	86
6.5.2	Evaluating the necessity of importance ranking	87
6.5.3	Entity Attack	87
6.5.4	Results of human evaluation	88
6.6	Discussion and limitations	89
6.7	Conclusions.	90
7	Fuzzy representation of discontinuous entities	91
7.1	Introduction	92
7.2	Related Work	93
7.3	Methods	94
7.3.1	The FuzzyBIO representation scheme	94
7.3.2	Named entity recognition of ADEs	94
7.3.3	Concept normalization of ADEs	94
7.3.4	Evaluation	94
7.4	Data.	95
7.5	Results	95
7.5.1	Intrinsic evaluation	95
7.5.2	Extrinsic evaluation	96
7.6	Discussion	96
7.7	Conclusion	97
III	Extracting coping strategies for adverse drug effects	99
8	The discovery of recommended coping mechanisms	101
8.1	Related work	103
8.2	Data	105
8.2.1	Data collection	105
8.2.2	Data annotation	106
8.2.3	Coping Strategy Ontology	108
8.2.4	Adding negative examples	109
8.3	Methods	111
8.3.1	Data preprocessing	111
8.3.2	ADE extraction and data selection	111
8.3.3	Coping Strategy extraction	112
8.3.4	Negation detection.	113

8.3.5	Relation extraction	114
8.3.6	Data post-processing	114
8.4	Results	114
8.4.1	Data description	114
8.4.2	Named entity recognition	115
8.4.3	End-to-end extraction	116
8.4.4	Case study on GIST ADE coping	117
8.5	Discussion	121
8.5.1	Comparison of approaches	121
8.5.2	Relevance of our findings	121
8.5.3	Potential application settings	124
8.5.4	Limitations.	125
8.5.5	Future work	125
8.6	Conclusion	126
IV	GIST as a case study	127
9	Patient forums as a complementary data source	129
9.1	Introduction	130
9.2	Materials and methods	131
9.2.1	Data collection.	131
9.2.2	Machine learning pipeline	131
9.2.3	Data analysis.	133
9.3	Results	134
9.4	Discussion	138
9.5	Conclusion	140
10	Comparing questionnaire and forum data	141
10.1	Introduction	143
10.2	Methods	144
10.2.1	Study design and participants	144
10.2.2	Recruitment and data collection	144
10.2.3	Study measures	145
10.2.4	Statistical analysis	145
10.3	Results	146
10.3.1	Participants	146
10.3.2	Prevalence scores	146
10.3.3	Relation between questionnaire and forum symptoms	146
10.4	Discussion	149
11	Assessing sample bias	153
11.1	Introduction	154
11.2	Methods	156
11.2.1	Study design & participants	156
11.2.2	Survey	156
11.2.3	Data analysis.	158

11.3 Results	158
11.3.1 Participants	158
11.3.2 Social media usage.	160
11.3.3 Reasons for abstaining from online communication with peers	160
11.3.4 Reasons for engaging with patient forums	161
11.3.5 Characteristics of the patient forum users	161
11.4 Discussion	165
11.4.1 Summary of findings.	165
11.4.2 Comparison with existing literature	165
11.4.3 Limitations.	166
11.4.4 Future work and recommendations	167
11.5 Conclusion	168
V Discussion	169
12 Discussion	171
12.1 Main findings	171
12.2 Answer to main research question	181
12.3 Future research	182
12.3.1 Mining experiential knowledge from social media	182
12.3.2 Ontology mapping and interoperability	183
12.3.3 Dealing with real-world data	185
12.4 Recommendations	187
12.4.1 Knowledge discovery from social media	187
12.4.2 Privacy and adopting FAIR metadata standards	188
12.4.3 Developing annotation guidelines	189
12.4.4 Long-term integration into healthcare	191
References	193
Summary	229
Samenvatting	231
Acknowledgements	233
Appendices	235
A Technical details of ADE extraction	237
A.0.1 Data preprocessing	237
A.0.2 Extracting ADEs from text	238
A.0.3 ADE normalization	239
A.0.4 Linking ADEs to medication	240

B Supplementary files for Chapter 8	249
C Supplementary Tables for Chapter 10	261
Curriculum Vitæ	267
List of Publications	269
SIKS Dissertation Series	271