Leveraging Social Media as a Source for Clinical Guidelines: A Demarcation of Experiential Knowledge

Jia-Zhen Chan j.m.chan@student.vu.nl
Florian Kunneman {f.a.kunneman,r.morantevallejo}@vu.nl
Roser Morante

Lea Lösch and **Teun Zuiderent-Jerak**
{lea.loesch,teun.zuiderent-jerak}@vu.nl
Vrije Universiteit Amsterdam

Abstract

In this paper we present a procedure to extract posts that contain experiential knowledge from Facebook discussions in Dutch, using automated filtering, manual annotations and machine learning. We define guidelines to annotate experiential knowledge and test them on a subset of the data. After several rounds of (re-)annotations, we come to an inter-annotator agreement of $K = 0.69$, which reflects the difficulty of the task. We subsequently discuss inclusion and exclusion criteria to cope with the diversity of manifestations of experiential knowledge relevant to guideline development.

1 Introduction

Messages shared on social media platforms are widely acknowledged as a source to gain insights into current events and the related vox populi. Recently, this source of information has been extensively leveraged for identifying topics of interest (Kellert and Mahmud Uz Zaman, 2022), emotions (Aduragba et al., 2022), and stances (Hossain et al., 2020; Glandt et al., 2021) during the COVID-19 pandemic. A lesser scrutinized type of information is experiential knowledge in the medical domain: messages that describe practical contact with diseases or treatments (e.g.: having been contaminated, having received a vaccination), which may include contextual information that is relevant to the experience (e.g.: being a doctor, having certain allergies). Such information is valuable for clinical guideline development, where social media may provide insights into latent experiences that are relevant to sharpen guidelines on, for example, exemptions and communication to patients.

In this paper we discuss the task of automatically identifying experiences in Facebook messages in Dutch. As an input for clinical guidelines, experiences on social media are most valuable when they are provided with context and accompanied with value considerations. Discussions following news reports on dedicated Facebook pages are a relevant source in that respect, since these messages tend to be more extended than, for example, Twitter messages, and may express personal disclosures as well. There are, however, two prominent challenges to identifying experiences from these discussions. First, only a small part of these messages share an experience, making it challenging to locate examples to train and test a machine learning classifier on. Second, experiences take a diversity of manifestations, posing difficulties to merely reach a sufficient agreement among humans. This is also reported in several studies that aim to identify experiences in Twitter (Sarker et al., 2020) and patient fora (Liu et al., 2015). In this paper we present our approach to these challenges, contributing an iterative procedure to zoom in on experiences within Dutch Facebook discussions on the topic of COVID-19 vaccinations, as well as an annotation guideline catered for annotating experiences as input for vaccination guidelines.

2 Related work

The most common aim for targeting experiential information from social media messages is to inquire into unforeseen effects of medicine intake (Alvaro et al., 2015; Liu et al., 2015; Cavazos-Rehg et al., 2016; Klein et al., 2017; Oostdijk et al., 2019; Kim et al., 2020; Sarker et al., 2020). Other aims are to gain insight into the experiences of particular patient groups (Stemmer et al., 2021) and to inquire into topics of discussions related to HPV vaccinations, where the sharing of experiences is identified as one of these topics (Surian et al., 2016). Studies that aim at identifying experiences mostly focus on Twitter as a data source, while user fora are leveraged to a lesser extent (Liu et al., 2015; Oostdijk et al., 2019). In contrast to these studies, we set out to identify experiences as input for rapid medical guideline development, and make use of Facebook discussions as a data source. While ex-

Proceedings of the 29th International Conference on Computational Linguistic, pages 203–208
October 12–17, 2022.

periences shared on social media have not been used before as evidence for clinical guidelines, automated approaches have been applied to swiftly identify scientific papers on a particular medical context to facilitate rapid clinical guideline development (Whittington et al., 2019).

In alignment with most of the studies that target experiences, we deployed manual annotations to come to a ground truth and encountered difficulties to reach a sufficient inter-annotator agreement on this task. Klein et al. (2017) ascribe these difficulties to the wide range of linguistic patterns by which experiences are manifested. We follow their solution to develop an annotation guideline that includes the different variants. In line with Alvaro et al. (2015), we discuss the particular dilemma's that pose difficulties to decide on an annotation.

A hampering factor to our endeavour to gain insight into experiences related to COVID-19 vaccinations, is that the topic of vaccination is highly debated and therefore dominated by opinionated posts rather than experiences. Participants in the debate express strong attitudes in favor and against vaccines, discussing issues related to the safety of vaccinations, the side effects, and the moral aspects of enforcing mandatory vaccination (Wolfe and Sharp, 2002; Mollema et al., 2015). Nowadays, with collaborative media, anyone can join in a discussion and share information and opinions. This makes it difficult to attest the reliability of on-line content (Zummo, 2017). Doubts about the reliability of vaccines can easily grow among the population under the influence of negative information about vaccines or unbalanced reports of vaccine risk (Betsch and Sachse, 2012).

Knowing what people think about vaccines and why people decide to vaccinate or not has always been interesting for health practitioners who need to be adequately informed on public perception of the safety and necessity of vaccines. There have been efforts aimed at annotating data about the vaccination debate (Morante et al., 2020; Torsi and Morante, 2018) and at automatically finding opinions in favour or against vaccines (Kunneman et al., 2020; Chen and Crooks, 2022; Cascini et al., 2022).

3 Identifying experiences from Facebook discussions

Our study sets off from a set of Facebook posts discussing news articles related to COVID-19 vaccinations, without a clear notion of the frequency and characteristics of messages that share an experience in these discussions. Next we describe the iterative procedure that we conducted to identify experiential messages and collect enough examples to train a machine learning classifier on. An overview of the stages of this procedure and samples used is presented in Table 1.[1]

3.1 Data

In total we collected 230,863 Facebook comments on news articles about COVID-19 vaccination posted by the four Dutch news outlets (NOS, NU.nl, Telegraaf and NRC) on their public Facebook pages between 01.12.2020 and 08.06.2021. Relevant articles from the respective news outlets were located using the following query terms "(Corona* OR COVID* OR COVID-19) AND (vaccinatie* OR vaccin OR inenten [Dutch for 'vaccinate'] OR prik [Dutch for 'shot'])". Using the Facebook Pages API[2], the post ID, the comment text, the posting date and the number of likes and comments were retrieved.

3.2 Initial data annotation

We developed an initial three stage rudimentary filter to locate the parts of the data that are likely to contain experiences. First, considering that experiences tend to be lengthy to describe, only comments exceeding 250 characters were kept. Second and third, accounts of experiences tend to show a higher degree of subjectivity and sentiment polarity in their wording. In order to select comments with a higher degree of subjectivity and sentiment, a sentiment analysis was carried out with the python package Pattern (De Smedt and Daelemans, 2012).[3] All comments were thereby assigned a sentiment value between -1 (negative) and +1 (positive) and a value between 0 (objective) and 1 (subjective) indicating their degree of subjectivity. Comments were retained if they had an above-average subjectivity score ($\geq$ 0.4) and a sentiment score of $\leq$ -0.25 or $\geq$ 0.25.

The initial filtering step resulted in a set of 5,702 comments. A random sample of 500 comments was then coded independently by two researchers on the presence of experiential knowledge. The definition of experience-based knowledge related

[1] Ids of the messages and individual annotations will be made available.

[2] https://developers.facebook.com/docs/pages/

[3] https://github.com/clips/pattern

Data source	Size	Part of	Criteria	IAA
Complete	230,863			
Filter 1	5,702	Complete	Message size	
			High subjectivity	
Sample 1	500	Filter 1		0.66
Filter 2	119	Filter 1	'My' + [family member]	0.60
Training	70,830	Complete		
Test	49,034	Complete	Disjoint from training	
Filter 3	1,258	Training	'My' + [family member]	
			'have...had'	
			hypernym to 'feel'	
Sample 2	500	Test	> 0.90 classifier confidence (250 posts)	0.59
			$> 0.50 < 0.90$ classifier confidence (250 posts)	
Sample 3	500	Test	> 0.90 classifier confidence (250 posts)	0.56^{i}
			$> 0.50 < 0.90$ classifier confidence (250 posts)	0.69^{ii}

Table 1: Overview of the samples that were drawn in the process of identifying experiential messages from Facebook discussions on COVID-19 vaccinations.Inter-annotator agreement (IAA) is measured in Cohen's Kappa. ^{i} First round. ^{ii}Second round.

to COVID-19 vaccination in comments was kept fairly broad, including both first-hand and second-hand experience. Our understanding is closest to the definition of "experience" from The Oxford Pocket Dictionary of Current English: "practical contact with and observation of facts or events". Whereas multiple experiences could be mentioned in a single post, the annotation target was only to assess whether at least one experience was mentioned.

The annotation effort indicated that 10% of these comments contained experiential knowledge, at a moderate inter-annotator agreement (0.66 Cohen's Kappa). We analysed the comments coded as "experience" to find specific features characteristic of these comments. The observation that these descriptions often contained references to a close contact was incorporated into the filter by additionally filtering for comments that included the combination of the words "my" and a close contact, such as a family member, e.g. "my grandma". This search resulted in 119 comments of which about 77% indicated experiential knowledge, which was confirmed by another round of annotation by the same two annotators who reached a moderate inter-annotator agreement (0.60 Cohen's Kappa).

To obtain a higher number of posts to train a machine learning classifier on, we expanded the latter filter rule with two additional filter rules and applied the filter to a sample of 70,830 Facebook posts (not complying with the initial filter based on length, subjectivity and sentiment). The additional filter rules were 1) including the pattern 'have ... had' (targeting expressions of having had COVID-19 or another disease) and 2) including a first- or second-level hypernym of the verb 'feel' (from the Open Dutch Wordnet (Postma et al., 2016)). This resulted in 1,258 (distantly supervised) examples of experiences as input for machine learning.

3.3 Experience modelling

We trained a machine learning classifier on the training set featuring 70,830 messages of which 1,258 were labeled as experience in a distantly supervised way. The performance of different algorithms and feature weightings was then compared by applying them to the 500 annotated posts in the previous phase, using Precision, Recall, F1 and Area under the ROC curve as evaluation metrics.

As preprocessing, the posts were cleaned of URLs, emojis, punctuation, numbers and symbols, which we did not consider predictive for identifying formulations of experiences, and the remaining tokens were lowercased. The cleaned and normalized texts were tokenized and lemmatized by means of the Dutch Stanza pipeline (Qi et al., 2020). We experimented with two different algorithms, Logistic Regression and XGBoost, which are common but distinct approaches to supervised machine learning, have yielded good results on several NLP tasks and lead to interpretable models. We also compared to feature weightings, tf*idf and binary. As a baseline we applied the three filters in a rule-based manner

System	P	R	F1	AUC
Baseline	0.32	0.14	0.20	
XGBoost (binary weighting)	0.39	0.59	0.47	0.74
XGBoost (tf*idf weighting)	0.43	0.53	0.47	0.73
Logistic Regression (binary weighting)	0.37	0.20	0.26	0.58
Logistic Regression (tf*idf weighting)	0.38	0.45	0.41	0.68

Table 2: Results on classifying Facebook posts sharing an experience in an annotated sample of 500 posts with a support of 49 posts labeled as experience (P=Precision, R=Recall, AUC=Area under the ROC Curve).

– if any of the given patterns matched the text in a post, the post was labeled as experience.

The outcomes of the machine learning experiment are presented in Table 2. The optimal machine learning set-up was an XGBoost classifier using a binary weighted lemma's representation, yielding an F1-score of 0.47 on predicting experiences in the annotated sample of 500 comments, considerably outperforming a rule-based filter (F1-score of 0.20). The significant difference in recall (0.59 vs. 0.14) is showing the generalisability of the machine learning approach. Note that the test set was considerably skewed with only 10% of the messages sharing an experience.

The best-performing classifier was subsequently used to identify experiences in a held-out set of 49,034 posts. From these classified posts, we extracted two samples to manually annotate for the number of experiences: a sample of 250 posts that were classified at a classifier confidence above 0.90, and a sample of 250 posts classified at a confidence between 0.50 and 0.90. This enabled us to inquire into the utility of the model in relation to classifier confidence when applied to unseen data. Based on the definition of experience used in the first annotation round the posts were coded by four annotators, who reached a slightly weak agreement (0.59 Cohen's Kappa).

To come to a higher agreement rate, the disagreements were discussed among the annotators and a set of inclusion and exclusion criteria were formulated in a more extensive annotation guideline. Three of the annotators then annotated an additional sample of again 500 posts (250 times at a classifier confidence above 0.90, 250 times between 0.50 and 0.90), surprisingly reaching an agreement of 0.56. They again discussed the disagreements, finding that some of the criteria in the annotation guidelines were leading to more confusion. Based on this discussion the guidelines were adapted, and the same set of posts were again annotated by the group of three. There were approximately 3 weeks in between the first and second round on these posts. A moderate agreement of 0.69 was reached on this set.

4 Annotation guideline and development

The annotation guideline including different manifestations of experiences and examples was written after the annotations of Sample 2 and refined after the annotations of Sample 3.[4] For reasons of space, we focus here on the cases that posed most difficulties to decide on an annotation:

- **Indirectness of experience**. Posts that describe a second-hand experience ('My mother had...') are arguably just as relevant as personal experiences. There is, however, a point where the described experience is too general to remain trustworthy ('I know of people who...'). We decided to exclude the most generic phrasing of experiences, but some cases in the annotation set still posed doubts ('A friend of my mother...').
- **Non-experiences**. Posts that state a non-experience ('I did not receive an invitation yet') were deemed relevant as well, since they may convey valuable information.
- **Personal disclosure**. A person may share information on group membership (working in a hospital) or having certain feelings ('I feel pressure to make a decision'), which may be of relevance as guideline input. Such personal disclosures are not strictly an experience, but we chose to include it.
- **Routines**. Some posts share a person's routine ('visiting my mother every 1,5 weeks), which are not experiences from a bounded period in the past, but do convey information about events that take place and may be of relevance.
- **Sources of information**. In a discussion on the topic of vaccinations, many posts discuss sources of information and their reliability. In some cases, obtaining information from a doctor or medical specialist is mentioned, which could be seen as an experience. We

[4]The final version of the guideline can be downloaded from the appendix: https://surfdrive.surf.nl/files/index.php/s/r1UKVCWhAwVUo2P

chose to exclude this as it relates too much to sources of information.

5 Conclusion

We set out to identify experiential knowledge from Facebook discussions in Dutch in relation to COVID-19 vaccinations, using filtering criteria, manual annotations and machine learning in order to broaden the filter and draw more samples for manual annotation. After several annotation rounds, we defined annotation guidelines for experiential knowledge. Annotators reached an agreement of 0.69 Cohen's Kappa, which indicates that the task is rather subjective with diverse manifestations of the targeted type of posts.

Messages on social media provide information that is difficult to obtain in other ways for rapid guideline development, with access to a diversity of backgrounds and experiences. Before being able to integrate this type of information, insight is needed into what separates utterances baring experiential information from other types of messages in Facebook discussions. The current endeavour to filter for such messages and discuss their characteristics has been a crucial step towards this point, but needs to be further refined to come to a machine learning classifier that yields a sufficient performance to be integrated in a hybrid workflow where a clinical guideline expert can further inspect the collected messages. Future work to come to this point includes experimenting with different algorithms, feature representations and distant supervision filters, and further refining the annotation guideline.

References

Olanrewaju Tahir Aduragba, Jialin Yu, Alexandra I. Cristea, and Lei Shi. 2022. Detecting fine-grained emotions on social media during major disease outbreaks: Health and well-being before and during the covid-19 pandemic. In *AMIA Annu Symp Proc. 2021*, pages 187–0–196. AMIA.

Nestor Alvaro, Mike Conway, Son Doan, Christoph Lofi, John Overington, and Nigel Collier. 2015. Crowdsourcing twitter annotations to identify first-hand experiences of prescription drug use. *Journal of biomedical informatics*, 58:280–287.

C. Betsch and K. Sachse. 2012. Dr. jekyll or mr. hyde? how the internet influences vaccination decisions: recent evidence and tentative guidelines for online vaccine communication. *Primary Care: Clinics in Office Practice*, 30(25):3723–3726.

Fidelia Cascini, Ana Pantovic, Yazan A. Al-Ajlouni, Giovanna Failla, Valeria Puleo, Andriy Melnyk, Alberto Lontano, and Walter Ricciardi. 2022. Social media and attitudes towards a covid-19 vaccination: A systematic review of the literature. *eClinical Medicine*, 48(101454):3723–3726.

Patricia A Cavazos-Rehg, Shaina J Sowles, Melissa J Krauss, Vivian Agbonavbare, Richard Grucza, and Laura Bierut. 2016. A content analysis of tweets about high-potency marijuana. *Drug and alcohol dependence*, 166:100–108.

Qingqing Chen and Andrew Crooks. 2022. Analyzing the vaccination debate in social media data pre- and post-covid-19 pandemic. *International Journal of Applied Earth Observation and Geoinformation*, 110:102783.

Tom De Smedt and Walter Daelemans. 2012. " vreselijk mooi!"(terribly beautiful): A subjectivity lexicon for dutch adjectives. In *LREC*, pages 3568–3572.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611, Online. Association for Computational Linguistics.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Olga Kellert and Md Mahmud Uz Zaman. 2022. Using neural topic models to track context shifts of words: a case study of COVID-related terms before and after the lockdown in April 2020. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 131–139, Dublin, Ireland. Association for Computational Linguistics.

Myeong Gyu Kim, Jungu Kim, Su Cheol Kim, and Jaegwon Jeong. 2020. Twitter analysis of the nonmedical use and side effects of methylphenidate: machine learning study. *Journal of medical Internet research*, 22(2):e16466.

Ari Klein, Abeed Sarker, Masoud Rouhizadeh, Karen O'Connor, and Graciela Gonzalez. 2017. Detecting personal medication intake in twitter: an annotated corpus and baseline classification system. In *BioNLP 2017*, pages 136–142.

Florian Kunneman, Mattijs Lambooij, Albert Wong, Antal van den Bosch, and Liesbeth Mollema. 2020. Monitoring stance towards vaccination in twitter messages. *BMC medical informatics and decision making*, 20(1):1–14.

Yunzhong Liu, Yi Chen, Jiliang Tang, and Huan Liu. 2015. Context-aware experience extraction from online health forums. In *2015 International Conference on Healthcare Informatics*, pages 42–47. IEEE.

Liesbeth Mollema, Irene Anhai Harmsen, Emma Broekhuizen, Rutger Clijnk, Hester De Melker, Theo Paulussen, Gerjo Kok, Robert Ruiter, and Enny Das. 2015. Disease detection or public opinion reflection? Content analysis of tweets, other social media, and online newspapers during the measles outbreak in The Netherlands in 2013. *Journal of medical Internet research*, 17(5).

Roser Morante, Chantal van Son, Isa Maks, and Piek Vossen. 2020. Annotating perspectives on vaccination. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4964–4973, Marseille, France. European Language Resources Association.

Nelleke HJ Oostdijk, Mattijs S Lambooij, Peter Beinema, Albert Wong, Florian A Kunneman, and Peter HJ Keizers. 2019. Fora fuelling the discovery of fortified dietary supplements–an exploratory study directed at monitoring the internet for contaminated food supplements based on the reported effects of their users. *Plos one*, 14(5):e0215858.

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open Dutch WordNet. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 302–310, Bucharest, Romania. Global Wordnet Association.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Abeed Sarker, Sahithi Lakamana, Whitney Hogg-Bremer, Angel Xie, Mohammed Ali Al-Garadi, and Yuan-Chi Yang. 2020. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. *Journal of the American Medical Informatics Association*, 27(8):1310–1315.

Maya Stemmer, Yisrael Parmet, and Gilad Ravid. 2021. What are ibd patients talking about on twitter? In *International Conference on ICT for Health, Accessibility and Wellbeing*, pages 206–220. Springer.

Didi Surian, Dat Quoc Nguyen, Georgina Kennedy, Mark Johnson, Enrico Coiera, Adam G Dunn, et al. 2016. Characterizing twitter discussions about hpv vaccines using topic modeling and community detection. *Journal of medical Internet research*, 18(8):e6045.

Benedetta Torsi and Roser Morante. 2018. Annotating claims in the vaccination debate. In *Proceedings of the 5th Workshop on Argument Mining*, pages 47–56, Brussels, Belgium. Association for Computational Linguistics.

Craig Whittington, Todd Feinman, Sandra Zelman Lewis, Greg Lieberman, and Michael del Aguila. 2019. Clinical practice guidelines: Machine learning and natural language processing for automating the rapid identification and annotation of new evidence.

Robert M Wolfe and Lisa K Sharp. 2002. Anti-vaccinationists past and present. *BMJ: British Medical Journal*, 325(7361):430.

Marianna Lya Zummo. 2017. A linguistic analysis of the online debate on vaccines and use of fora as information stations and confirmation niche. *International Journal of Society, Culture & Language*, 5(1):44.