Automatic Identification of Ruptures in Transcribed Psychotherapy Sessions

Adam Tsakalidis[1,2], **Dana Atzil-Slonim**[3], **Asaf Polakovski**, **Natalie Shapira**[3], **Rivka Tuval-Mashiach**[3], **Maria Liakata**[1,2,4]

[1] Queen Mary University of London, London, United Kingdom
[2] The Alan Turing Institute, London, United Kingdom
[3] Bar Ilan University, Ramat Gan, Israel
[4] University of Warwick, Coventry, United Kingdom
a.tsakalidis@qmul.ac.uk

Abstract

We present the first work on automatically capturing alliance rupture in transcribed therapy sessions, trained on the text and self-reported rupture scores from both therapists and clients. Our NLP baseline outperforms a strong majority baseline by a large margin and captures client reported ruptures unidentified by therapists in 40% of such cases.

1 Introduction

The client-therapist relationship within a psychotherapy treatment ('therapeutic alliance') is considered a powerful predictor of therapy success across treatment modalities and disorders (Flückiger et al., 2018; Norcross and Lambert, 2019). Conversely, when a tension or a breakdown (*rupture*) occurs in the therapeutic alliance, it can often lead to unilateral termination of the treatment by the client or to poor psychotherapy outcomes (Eubanks et al., 2018). However, when alliance ruptures are recognised they can become meaningful therapeutic events (Chen et al., 2018). Indeed, alliance ruptures have been found to be beneficial to the therapeutic process and outcome when they are recognized and followed by repair of the rupture (Stevens et al., 2007; Stiles et al., 2004) and to hinder the process or outcome of therapy when they go unrecognized (Chen et al., 2018).

Challenges in capturing alliance rupture: Most studies have explored alliance ruptures using self-reports at relatively low time resolution (once each session, typically weekly). However, ruptures may occur at higher time resolutions within a session (Coutinho et al., 2014). In addition, standardized subjective measures have critical shortcomings, including the extent of participants' self-insights, willingness to complete questionnaires, and the restricted choice of responses (Kazdin, 2016). Recent studies have used within-session coding tools to detect ruptures moment-by-moment during a session, yielding important insights into the within-session processes that lead to ruptures (e.g., (Eubanks et al., 2015). These insights have been used to train therapists to recognize ruptures when they happen (Eubanks-Carter et al., 2015). However, since observational human-coding is very labor intensive and expensive, these studies have focused on a small number of therapeutic components in a small sample of clients and at limited time points.

Benefits of capturing alliance rupture from text originating from the transcribed dialogue between therapist and client during therapy sessions include:

- Detecting alliance rupture even when therapists or clients are unaware of it. This would allow signaling the rupture to therapists and help them acknowledge it. Such information may be used alongside existing monitoring tools to inform therapists about meaningful instances of alliance rupture that went unrecognized.
- Subtler and more implicit content associated with a rupture would be captured, increasing our understanding of the specific moments and reasons for it.
- Alliance rupture would be captured in a cost-effective manner.

Contributions: To the best of our knowledge there is no work on capturing alliance rupture automatically from transcribed therapist or client utterances. Recently Goldberg et al. (2020) used 1,235 transcribed recorded sessions with client reported alliance to automatically predict per session alliance using the text from both therapist and client. They used four variants of a linear regression model with linguistic features from either the therapist or client. Their best performing model was only 0.02 more accurate than a baseline predicting the average alliance rating. They also provided a list of unigrams which correlate most with high and low alliance scores respectively.

Here we make the following contributions:

- We present the first work on automatically capturing alliance rupture (rather than alliance) trained on transcribed therapy sessions and self-reported rupture scores.
- We provide a detailed description of the dataset creation.
- We provide strong NLP baselines which outperform majority baselines by a large margin. Moreover we have an original privacy preservation setting whereby the data given to the NLP researchers was in encrypted format, facilitating the collaboration of NLP researchers with clinicians and companies with strong privacy concerns.
- We provide a qualitative analysis of examples where our NLP baselines capture client reported ruptures unrecognised by the therapist.

2 Dataset Description

Clients: were sampled from a pool of clients receiving individual psychotherapy at a university training outpatient clinic. Data were collected between Aug'14-Aug'16 as part of the clinic's regular practice of monitoring clients' progress. From an initial sample of 180 consented clients 34 (18.9%) dropped out. Clients were selected according to two criteria: (a) treatment duration of at least 15 sessions and (b) availability of full data, including audio recordings and session-by-session questionnaires. Clients were also excluded based on the M.I.N.I. 6.0 (Sheehan et al., 1998) if they were diagnosed as severely disturbed. The data of 68 (37.8%) clients who met the inclusion criteria were transcribed, for a total of 873 transcribed sessions. Clients were above the age of 18 (μ_{age}=39.06, SD=13.67, range 20–77), the majority of whom were women (58.9%). 53.5% had at least a bachelor's degree, 53.5% reported being single, 8.9% were in a committed relationship, 23.2% were married and 14.2% were divorced or widowed. Clients' diagnoses were established based on the Mini International Neuropsychiatric Diagnostic Interview for Axis I DSM-IV diagnoses (Sheehan et al., 1998). 22.9% of the clients had a single diagnosis, 20.0% had two and 25.7% had three or more. The most common diagnoses were comorbid anxiety and affective disorders (25.7%), followed by other comorbid disorders (17.1%), anxiety disorders (14.3%), and affective disorders (5.7%).

Therapists and Therapy: Clients were treated by 52 therapists at various stages of their clinical training. Clients were assigned to therapists in an ecologically valid manner based on therapist availability and caseload. 42 therapists treated one client each; eight treated two clients. Each therapist received one hour of individual supervision biweekly and four hours of group supervision on a weekly basis. All therapy sessions were audiotaped for supervision by senior clinicians. Supervision focused heavily on reviewing audiotaped case material and technical interventions designed to facilitate the appropriate use of therapist interventions. Individual psychotherapy consisted of once- or twice-weekly sessions. The language of therapy was Modern Hebrew (MH). The dominant approach in the clinic includes a short-term psychodynamic psychotherapy treatment model (e.g., (Blagys and Hilsenroth, 2000; Shedler, 2010; Summers and Barber, 2009). On average, treatment length was 37 sessions (SD=23.99, range=18–157). Treatment was open- ended in length, but given that psychotherapy was provided by clinical trainees at a university-based outpatient community clinic, treatment duration was often restricted to 9 months.

Instruments and Procedure: Clients and/or therapists responded to several scales during the treatment, including the Outcome Rating Scale (Miller et al., 2003) and the Post-Session Questionnaire (PSQ). In this work, we focus specifically on the *alliance ruptures*. Alliance ruptures were assessed after each session with one question to the therapist and client: "Did you experience any tension, misunderstanding, conflict or disagreement in the relationship with your client/therapist?". This item is answered subjectively on a 5-point Likert scale from 1 ('not at all') to 5 ('constantly') by the two involved entities separately. Following (Muran et al., 2009), a rupture was defined as any rating higher than 1 on the scale. The PSQ has been widely used in psychotherapy research and demonstrates sound psychometric properties, including predictive validity with a variety of process indices such as the Working Alliance Inventory (Tracey and Kokotovic, 1989). Here the PSQ mean score was 2.06 (SD=1.43).

Transcription: Due to the high associated cost manual transcriptions were conducted alternately (sessions 2, 4, 6, etc.). In cases where material was incomplete (e.g., questionnaire or poor recording quality), the following session was transcribed in-

stead. The transcriber team was composed of seven graduate students in the University's psychology department. The transcribers went through a one day training workshop which included how to handle private/sensitive information; monthly meetings were held throughout the transcription process to supervise the quality of their work. The transcription protocol followed general guidelines, as described in (Mergenthaler and Stinson, 1992; Albert et al., 2013). The word forms, the form of commentaries and the use of punctuation were kept as close as possible to the speech presentation. Everything was transcribed, including word fragments as well as syllables or fillers (e.g., "ums", "ahs", "you know"). The transcripts included elisions, mispronunciations, slang, grammatical errors, non-verbal sounds (e.g., laughs, cry, sighs), and background noises. The rules were limited in number and simple and the format used several symbols to indicate comments (e.g. '[...]' to indicate the correct form when the actual utterance was mispronounced).

There were 873 transcripts in total (the mean transcribed sessions per client was 12.56; SD=4.93). The transcriptions totaled about four million words over 150,000 talk turns (i.e., switching between speakers). On average, there were 5800 words in a session, of which 4538 (78%; SD=1409.62; range 416-8176) were client utterances and 1266 (22%; SD=674.99; range 160-6048) were therapist utterances.

Text Processing & Privacy: In morphologically rich languages such as Hebrew, each token may have multiple different morphological analyses where only one is pertinent to the context. To tackle this, we used the YAP parser (More and Tsarfaty, 2016), which performs a lexicon-based morphological analysis followed by joint morpho-syntactic disambiguation and dependency parsing. Finally, to work in a privacy preserving manner due to the sensitive nature of our data, we replaced each word with a token ID. We further used a separate mapping of the token IDs to indices in a dictionary of word vectors to share the data within our team for our experiments. The word vectors were also rotated, as an additional security step.

3 Experiments

Task: We define the problem of capturing rupture alliance as a binary classification task. In particular, we aim at identifying whether a rupture occurred within a session, given the language used by the therapist and/or the client during that session. The presence or absence of rupture is defined via the self-assessed questionnaire, which is completed by each of the client and therapist. We treat their responses as two separate tasks: (a) Client's Rupture (CR) prediction and (b) Therapist's Rupture (TR) prediction, where in each task the goal is to predict the corresponding self-reported outcome given the transcribed session as input.

Task	**Client's Rupture [CR]**			**Therapist's Rupture [TR]**		
Features	Client	Therapist	Both	Client	Therapist	Both
Majority	59.00	59.00	59.00	37.50	37.50	37.50
LogReg	**61.90**	61.30	58.80	45.60	46.60	**46.70**

Table 1: F-score for the two binary classification tasks.

Dataset: Since some of the transcriptions were not associated with alliance rupture labels, the final dataset used in our experiments consists of 849 transcribed sessions from 68 clients. Due to missing labels, the two tasks also have a different number of instances. There were 821 sessions for CR and 829 for the TR task. The distribution of the labels for the two tasks differs: for TR there is a balance between rupture *vs* no-rupture labels (48% *vs* 52%); the same does not hold for CR (23% *vs* 77%).

Experimental Setting: The input to our classifier in the text from a transcribed therapy session. We represent each session via dense word vectors consisting of: (a) the client's text, (b) the therapist's text and (c) both of them in concatenation. The vectors were obtained by training a skip-gram model (Mikolov et al., 2013) on a large collection of tweets in Hebrew. With each word represented as a 100-dim vector, we represent each session by averaging the dimensions of words used by either the client, therapist or both during the session.

We train a Logistic Regression for our two tasks, CR and TR. We perform a leave-one-client-out cross validation (68 folds) to avoid any potential bias in our evaluation (DeMasi et al., 2017; Tsakalidis et al., 2018; Harrigian et al., 2020). This way we can assess the model's ability to generalise in previously unseen clients. For each task, we experiment with the three types of representations discussed above. For evaluation we use the macro-averaged F-score between the two classes, averaged across all folds. We contrast performance against the majority (no-rupture) classifier to get some first insights into the difficulty of the tasks.

Results: Table 1 shows the macro-average F-score achieved in the two tasks, averaged across all

clients (folds). The performance on the CR task is higher compared to the TR task due to the imbalanced nature of our dataset. However, there is only a minor relative improvement of 4.9% in CR over the majority baseline (52.8% over a completely random classifier) compared to the 24.5% in TR. This large difference between the two tasks is attributed to the fact that therapists are trained to recognise ruptures and are more likely to report ruptures than miss a potential rupture. This makes the dataset more balanced in terms of rupture and non-rupture labels.

Next, we examine the performance on the 801 sessions were we have reports on rupture by both the therapist and the client.In particular, we are interested in inspecting cases of sessions where the client indicated that there was a rupture, but the therapist missed it. Therefore, we treat the label provided by the Client ('rupture') as our ground truth and test our models' performance based on them, when leveraging both of the Client's and the Therapist's text. Overall, there were 72 such cases (9%), as shown in Table 2. Logistic Regression trained for the TR task successfully identified 29 (40%) of these cases. This encouraging finding suggests that incorporating NLP methods for detecting such cases – which is of particular importance for therapists – could act as a tool too assist with rupture detection to improve psychotherapy treatment. On the other extreme combination of labelling shown in Table 2 (i.e., in 341 cases which both the client and the therapist reported as "no rupture"), there were 205 (60%) sessions that have been correctly classified by *both* of the CR and TR models jointly, while there were only 10 of these cases (3%) that were jointly misclassified by the two models. Overall, by considering only the rather "clear" 274 sessions (i) which have been given the same ground-truth label by both client and therapist and (ii) for which the CR/TR models agree on their prediction, the (%) macro-average F1-score is 70.9% (accuracy 83.6%). This suggests that the task of predicting rupture alliance by analysing the language used within a psychotherapy session is indeed feasible. However, there is plenty of room for improvement both in terms of language representation as well as modelling.

Finally, we inspect the language used within rupture *vs* non-rupture sessions. We are particularly interested the sessions that were labelled as 'rupture' by the client only (see Table 2) and also correctly identified by our model (40%). We find that most of them were withdrawal ruptures (see example in Table 3a). The literature on ruptures highlights two main subtypes: withdrawal and confrontation ruptures (Eubanks et al., 2018). In withdrawal ruptures (see example in Table 3b), the client moves away from the therapist and the work of therapy, e.g. by avoiding the therapist's questions or by hiding their dissatisfaction with therapy by being overly appeasing. In confrontation ruptures (see example in Table 3c), the client moves against the therapist by expressing anger or dissatisfaction with the therapist or treatment, or by trying to apply pressure on the therapist. It seems that it was easier for therapists in our sample to identify the occurrence of confrontational ruptures, which may be more apparent in the client's behavior than the withdrawal ruptures. The latter may be more subtle and less emotionally charged. This finding is in line with other qualitative studies showing that therapists tend to better recognize confrontational ruptures (Hill, 2010). It also highlights the importance of using automated methods to capture ruptures that are challenging for therapists to capture.

		Therapist	
		No rupture	Rupture
Client	No rupture	341	280
	Rupture	72	108

Table 2: Distribution of labels in the 801 sessions that were labelled by both entities (Therapist, Client).

4 Conclusion and Future Work

In this work we focused on the task of automatically predicting alliance rupture between a therapist and a client from the language of therapy sessions. We collected and transcribed sessions between clients and their therapists, conducted in Hebrew. We also obtained self-reported rupture labels for sessions by clients and therapists, used in clinical psychotherapy research. We tested baseline models leveraging the language used within a session to predict the occurrence of alliance rupture based on the perception of both the therapist and the client. We yield good performance and showcase the potential for using NLP for aiding therapists in identifying rupture during psychotherapy sessions. In the future we plan to build on our initial findings by incorporating contextual language models (Chriqui and Yahav, 2021; Devlin

I had to pick up my kind from his music lessons and I was busy and I asked my husband if he could take the child and he said he was busy and that I was the one who should give up.
Why do you think this is happening?
I always have to run from one thing to another. He's busy with his own affairs. But what did you ask? I'm not in focus.
No, no, it's okay, please continue.
I feel like I was unlucky in life. Yesterday I needed his help, but he is never there to help or hug me. I just don't have anyone who can do that for me. It's hard. I need someone who can support me. I never had such support in my life. I tried to get closer to him, but I feel that I am the only one whose needs are dismissed. He never gives up his needs. I feel so tired of all that. I have no desire to do anything.
We talked in the last session about your difficulties to bring your needs. But last time you also said that you felt closer to him, didn't you?
Yes, I should try to get closer to him, I don't know, maybe I am wrong.
How is it for you with other people?
I don't know.

(a) Example of part of a session that was labelled by the Client as 'rupture', but not from the Therapist. Logistic Regression trained for the TR task predicted that there is a 'rupture'.

I think I should be an employee instead of a boss. That pressure... I can't stand it. I'm not good at it. When a client comes I'm at the height of my enthusiasm, I have a lot of ideas on what to do, and I make plans & invest a lot of thought, I want it to be perfect, but something stops me, I cannot do it the way I want.
You are afraid of disappointing.
Yes, exactly. I invest too much time in planning and then something stops me from doing it. I want it to be perfect and I'm working on the planning and I'm getting exhausted. I feel so much pressure to implement the plan & then I just become lazy and unable to actually do it. Maybe if I was an employee then I would have cared less & the job would have been easier.
Sounds like there is a lot of pressure, also around the thought of finding another job.
No, it's not about finding another job.
But you also said - I feel that .. I have lots of strength and lots of motivation and I have many ideas, and suddenly when it comes to execution I can not find them.
There is some kind of fatigue, laziness, I feel I do not have the strength, not the physical strength, the mental strength.
Something stops you. Lets try to understand what it is.
I tend to postpone everything.
What do you postpone here?
Everything.
What do you postpone here, in treatment?
Nothing specific. I just tend to postpone everything.

(b) **Withdrawal rupture:** A translated snippet of a session where the client reported a 'rupture', but not the therapist. Logistic Regression trained for the TR task predicted that there is a 'rupture', agreeing with the client.

It's cold in here.
Cold?
Um .. this is, I'm coming here and the feelings are really .. confused, turbulent. I had a really completely confused week, I had a very very hard time at the end of the previous session.
Mmm..
It made me tense, and I was thinking if this form of treatment is good for me or if it's doing me any harm. I was looking for answers. I don't know if going deeper into things is good for me or if the right way for me is the opposite – to let go.
Mmmm
And I met again that person I have worked with last summer. He is helping me to raise my self-confidence. Sometimes that's what I need when I feel confused and unstable.
I hear you. I also thought a lot about the hard things you talked about in the previous meeting.
I felt overwhelmed and confused after the session.
Let's try to talk about what was it that you needed from me last time and that you felt that I did not provide.

(c) **Confrontational rupture:** This snippet of a translated transcribed session that was labelled by both client and therapist as a 'rupture'.

et al., 2019) and by developing models that can perform this task in a sequential and temporally sensitive manner. Finally, a limitation of our work stems from the fact that the clients and therapists come from a coherent background (linguistic, cultural). Confirming our findings via analysing data from therapy sessions across different backgrounds is an important future direction.

Acknowledgements

This work was supported by a UKRI/EPSRC Turing AI Fellowship to Maria Liakata (grant EP/V030302/1) and the The Alan Turing Institute (grant EP/N510129/1).

References

Aviad Albert, Brian MacWhinney, Bracha Nir, and Shuly Wintner. 2013. The hebrew childes corpus: transcription and morphological analysis. *Language resources and evaluation*, 47(4):973–1005.

Matthew D Blagys and Mark J Hilsenroth. 2000. Distinctive features of short-term psychodynamic-interpersonal psychotherapy: A review of the comparative psychotherapy process literature. *Clinical psychology: Science and practice*, 7(2):167–188.

Roei Chen, Dana Atzil-Slonim, Eran Bar-Kalifa, Ilanit Hasson-Ohayon, and Eshkol Refaeli. 2018. Therapists' recognition of alliance ruptures as a moderator of change in alliance and symptoms. *Psychotherapy Research*, 28(4):560–570.

Avihay Chriqui and Inbal Yahav. 2021. Hebert & hebemo: a hebrew bert model and a tool for polarity analysis and emotion recognition. *arXiv preprint arXiv:2102.01909*.

Joana Coutinho, Eugénia Ribeiro, Catarina Fernandes, Inês Sousa, and Jeremy D Safran. 2014. The development of the therapeutic alliance and the emergence of alliance ruptures.[el desarrollo de la alianza terapéutica y la aparición de rupturas en la alianza]. *Anales de Psicología/Annals of Psychology*, 30(3):985–994.

Orianna DeMasi, Konrad Kording, and Benjamin Recht. 2017. Meaningless comparisons lead to false optimism in medical machine learning. *PloS one*, 12(9):e0184604.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Catherine F Eubanks, J Christopher Muran, and Jeremy D Safran. 2015. Rupture resolution rating system (3rs): Manual. *Unpublished manuscript, Mount Sinai-Beth Israel Medical Center, New York*.

Catherine F Eubanks, J Christopher Muran, and Jeremy D Safran. 2018. Alliance rupture repair: A meta-analysis. *Psychotherapy*, 55(4):508.

Catherine Eubanks-Carter, J Christopher Muran, and Jeremy D Safran. 2015. Alliance-focused training. *Psychotherapy*, 52(2):169.

Christoph Flückiger, AC Del Re, Bruce E Wampold, and Adam O Horvath. 2018. The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy*, 55(4):316.

Simon B Goldberg, Nikolaos Flemotomos, Victor R Martinez, Michael J Tanana, Patty B Kuo, Brian T Pace, Jennifer L Villatte, Panayiotis G Georgiou, Jake Van Epps, Zac E Imel, et al. 2020. Machine learning and natural language processing in psychotherapy research: Alliance as example use case. *Journal of counseling psychology*, 67(4):438.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3774–3788.

Clara E Hill. 2010. Qualitative studies of negative experiences in psychotherapy.

Alan E Kazdin. 2016. *Methodological issues and strategies in clinical research*. American Psychological Association.

Erhard Mergenthaler and Charles Stinson. 1992. Psychotherapy transcription standards. *Psychotherapy research*, 2(2):125–142.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pages 3111–3119.

Scott D Miller, BL Duncan, J Brown, JA Sparks, and DA Claud. 2003. The outcome rating scale: A preliminary study of the reliability, validity, and feasibility of a brief visual analog measure. *Journal of brief Therapy*, 2(2):91–100.

Amir More and Reut Tsarfaty. 2016. Data-driven morphological analysis and disambiguation for morphologically rich languages and universal dependencies. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 337–348.

J Christopher Muran, Jeremy D Safran, Bernard S Gorman, Lisa Wallner Samstag, Catherine Eubanks-Carter, and Arnold Winston. 2009. The relationship of early alliance ruptures and their resolution to process and outcome in three time-limited psychotherapies for personality disorders. *Psychotherapy: Theory, Research, Practice, Training*, 46(2):233.

John C Norcross and Michael J Lambert. 2019. *Psychotherapy relationships that work: Volume 1: Evidence-based therapist contributions*. Oxford University Press.

Jonathan Shedler. 2010. The efficacy of psychodynamic psychotherapy. *American psychologist*, 65(2):98.

David V Sheehan, Yves Lecrubier, K Harnett Sheehan, Patricia Amorim, Juris Janavs, Emmanuelle Weiller, Thierry Hergueta, Roxy Baker, and Geoffrey C Dunbar. 1998. The mini-international neuropsychiatric interview (mini): the development and validation of a structured diagnostic psychiatric interview for

dsm-iv and icd-10. *The Journal of clinical psychiatry*.

Christopher L Stevens, J Christopher Muran, Jeremy D Safran, Bernard S Gorman, and Arnold Winston. 2007. Levels and patterns of the therapeutic alliance in brief psychotherapy. *American journal of psychotherapy*, 61(2):109–129.

William B Stiles, Meredith J Glick, Katerine Osatuke, Gillian E Hardy, David A Shapiro, Roxane Agnew-Davies, Anne Rees, and Michael Barkham. 2004. Patterns of alliance development and the rupture-repair hypothesis: Are productive relationships u-shaped or v-shaped? *Journal of Counseling Psychology*, 51(1):81.

R. F. Summers and J. P. Barber. 2009. *Psychodynamic therapy: A guide to evidence-based practice.* New York and London: Guilford Press.

Terence J Tracey and Anna M Kokotovic. 1989. Factor structure of the working alliance inventory. *Psychological Assessment: A journal of consulting and clinical psychology*, 1(3):207.

Adam Tsakalidis, Maria Liakata, Theo Damoulas, and Alexandra I Cristea. 2018. Can we assess mental health through social media and smart devices? addressing bias in methodology and evaluation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 407–423. Springer.