University of Southampton

UKRI Trustworthy Autonomous Systems Hub

University of Nottingham
UK | CHINA | MALAYSIA

SafeSpacesNLP

Exploring behaviour classification around online mental health conversations from a multi-disciplinary context - NLP, applied linguistics, social science and human-in-the-loop AI

UCREL CRS seminar

Dr Stuart E. Middleton[1], Dr Elena Nichele[2]
[1] University of Southampton, Electronics and Computer Science, sem03@soton.ac.uk
[2] University of Nottingham, School of Computer Science, elena.nichele@nottingham.ac.uk

5th May 2022

Overview

- Speakers

- SafeSpacesNLP Overview

- Current Research into Behaviour Classification and Human-in-the-loop NLP

- SafeSpacesNLP - Computer Science

- Challenges

- Current Research into Socio-Behaviour Analysis for Mental Health

- SafeSpacesNLP - Applied Linguistics / Social Science

- Challenges

- Summary

Speakers

- Dr Stuart E. Middleton

- Associate Professor, University of Southampton

- Research interests are Natural Language Processing (NLP), with a focus on Information Extraction (IE) and Socio-technical NLP within multi-disciplinary environments

- Active grants

  – PI - NERC-funded platform grant GloSAT NE/S015604/1

  – PI - UKRI TAS Hub funded agile project SafeSpacesNLP

  – CoI - ESRC-funded standard grant ProTechThem ES/V011278/1

Speakers

- Dr Elena Nichele

- Research Fellow, University of Nottingham

- Interested in communication and the impact language(s) can have on marketing (and vice versa) through technologies

- Research interests are cross-disciplinary

  – Computer-mediated communication

  – Business communication

  – Discourse analysis

  – Corpus analysis

SafeSpacesNLP Overview

- SafeSpacesNLP project

  - UKRI TASHub agile project; Sept 2021 to Aug 2022
    https://www.tas.ac.uk/safespacesnlp/

  - Behaviour classification NLP in a socio-technical NLP setting

  - Online harmful behaviours for children and young people

  - Kooth >> UK's leading mental health forum provider for young people; 200,000 post dataset

- Graph-based NLP >> text classification (behaviour markers) within conversations

- Human-in-the-loop NLP >> active learning & adversarial training

- Socio-behaviour analysis >> interviews, data coding & linguistic analysis, trust

- Multi-disciplinary teams

  - Computer Science, Criminology, Applied Linguistics

  - Stakeholder co-design and trust experiments

5

Current Research into Behaviour Classification and Human-in-the-loop NLP

- NLP events covering behaviour classification for online harms and mental health
  - Workshop on Online Abuse and Harms (WOAH)
    - Hate speech
    - https://www.workshopononlineabuse.com/
  - Workshop on Social Media Mining for Health Applications (#SMM4H)
    - Depression, Drug/medicine abuse, Medication
    - https://aclanthology.org/events/smm4h-2021/
  - Workshop on Computational Linguistics and Clinical Psychology (CLPsych)
    - Depression, Suicide, Self-harm, Gender identify, Anorexia etc.
    - https://clpsych.org/

Current Research into Behaviour Classification and Human-in-the-loop NLP

- NLP events covering behaviour classification for online harms and mental health

  - SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)

    - Hate speech

    - https://aclanthology.org/S19-2010/

  - SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text

    - Emotion detection

    - https://aclanthology.org/S19-2005/

7

Current Research into Behaviour Classification and Human-in-the-loop NLP

- Datasets - Hate Speech (mostly single posts no context)

  – [Waseem 2016] (16k tweets) >> sexism; racism

  – [Davidson 2017] (24k tweets) >> hate speech; offensive; neither

  – [Founta 2018] (80k tweets) >> abusive; hateful; spam; normal

  – [Basile 2019] HatEval - SemEval 2019 task 5 (12k tweets) >> hateful; not hateful

  – [Caselli 2020] OLID, AbuseEval (14k tweets) >> offensive; not offensive >> abusive; not abusive

  – [Vidgen 2020] East Asian Prejudice (20k tweets) >> hostility; criticism; counter speech; discussion of east Asian prejudice; non-related

  – [Wulczyn 2017] Wiki-detox (Wikipedia discussion comments, 100k manual label; 36M automated) >> personal attack; not personal attack

  – Kaggle toxic comment classification challenge dataset (159k Wikipedia comments) >> toxic; severe_toxic; obscene; threat; insult; identity_hate
  https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

Current Research into Behaviour Classification and Human-in-the-loop NLP

- Datasets - Hate Speech (mostly single posts no context)

  - [Davidson 2020] (5k reddit) >> 4 classes (hate group/person; hate idea/behaviour; hate way of communication; profanity)

  - [Rottger 2021] (3k hand crafted sents) >> hateful; non-hateful. 29 functional tests covering 11 classes of hate speech. For use as a testset for forensic analysis of models

  - [Price 2020] (44k comments on news stories) >> 6 hate classes (hostile, antagonistic, gross generalization, sarcastic, dismissive, condescending)

  - [Kurrek 2020] (40k reddit) >> 4 hate classes (derogatory, appropriative, non D&A, homonyms) and 12 subcategories

  - [Samghabadi 2020] (2k QA pairs) >> offensive; neural

  - [Van Hee 2018] (113k English, 78k dutch, AskFM QA pairs, 5k each cyberbullying) >> 4 roles (harasser, victim, bystander defender, bystander assistant) + 8 hate labels

Current Research into Behaviour Classification and Human-in-the-loop NLP

- Datasets - Emotion (mostly single posts no context)

  - [Rashkin 2019] Empathetic Dialogues Crowd-sourced role-played dialogue with 24,850 conversations, 2 speakers, 1 session, 32 conversation emotion labels

  - [Abdul-Mageed 2017] (1.6M tweet ID's + distant supervision emotion labels), 8/24 emotion labels as per [Putchick 1980]

  - [Sosea 2020] (8.5k sents), 8 emotion labels as per [Putchick 1980]

  - [Khanpour 2018] (2k forum/blog posts), 5 emotion labels as per [Putchick 1980]

Current Research into Behaviour Classification and Human-in-the-loop NLP

- Datasets - Mental Health (often **conversations** leading up to events)

  - [Inches 2012] PAN 2012 dataset online grooming (11k chat; 28k IRC; 268k omegle - ID's of predators listed) >> predator; not predator

  - [Pirina 2018] dataset depression (800 reddit posts) >> depression; no depression

  - [Zirikly 2019] CLPsych 2019 Shared Task (subreddit suicidewatch, 22k users, 50% from suicidewatch) >> no risk; low risk; moderate risk; severe risk http://users.umiacs.umd.edu/~resnik/umd_reddit_suicidality_dataset.html

  - [Park 2019] Korean async text chat Counselling sessions, 100 conversations (100 to 300 utterances per dialogue), 2 speakers, 4 classes (events; thoughts; emotions; behaviour) + tags = 40 unique labels

  - [Kayi 2018] dataset schizophrenia (373 paras written by confirmed patients); 400k tweet ID's from self-reported users

  - [Sharma 2020] Mental Health subreddits; TalkLife, 10k pairs (post, response) >> EPITOME coding scheme, 3 classes (emotion reaction; interpretation; explorations)
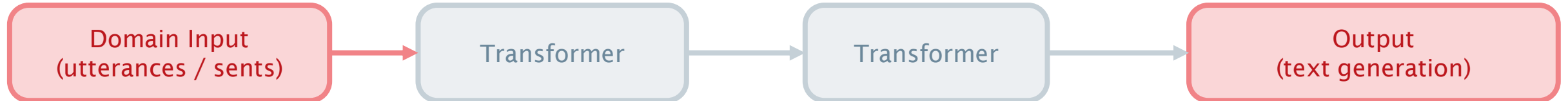
Current Research into Behaviour Classification and Human-in-the-loop NLP

- Short conversation classification

  – Text classification of single posts, 1 speaker

  – Dialogue classification of 2-3 conversation turns, 2 speakers

  – No temporal context

    • Complex behaviour and key moments not explicitly represented

- Long conversation classification

  – Dialogue classification of 100+ conversation turns, 2 speakers

  – Dialogue classification of 100+ conversation turns, 3+ speakers

  – Temporal context but currently not well studied [Xu 2021]

    • Most work concat conversation into big documents, Transformer has 512 token limit

    • Larger token windows (1000+) need larger numbers of examples (millions) to learn
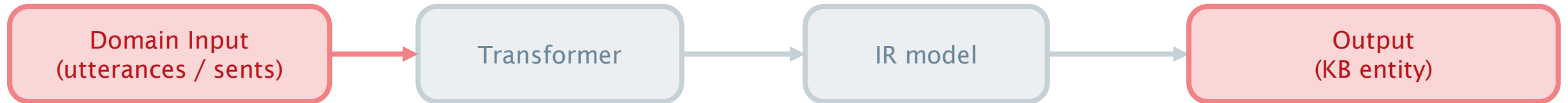
12

Current Research into Behaviour Classification and Human-in-the-loop NLP

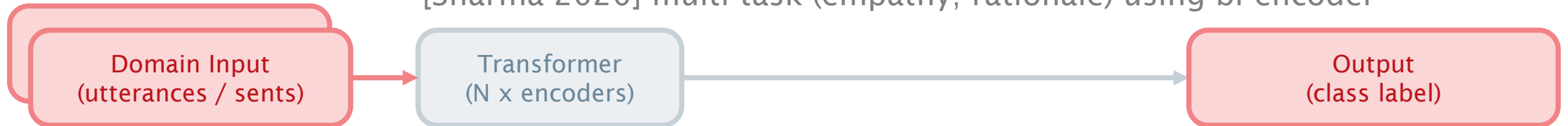- Long conversation NLP (augmented transformers, multi-task, bi-encoders)

[Xu 2021] augmented T1 = Sent Filter; Text Summarization

| Domain Input (utterances / sents) | → | Transformer | → | Transformer | → | Output (text generation) |

[Tigunova 2020] augmented T1 = keywords for searching Wikipedia

| Domain Input (utterances / sents) | → | Transformer | → | IR model | → | Output (KB entity) |

[Sharma 2020] multi-task (empathy; rationale) using bi-encoder

| Domain Input (utterances / sents) | → | Transformer (N x encoders) | → | Output (class label) |

University of
**Southampton**

Current Research into Behaviour Classification and Human-in-the-loop NLP

- Human-in-the-loop NLP

[Bunch 2020] [Middleton 2020] cyclic filtering/tuning for interactive sense making
[Hanafi 2020] explainable AI to understand/improve model
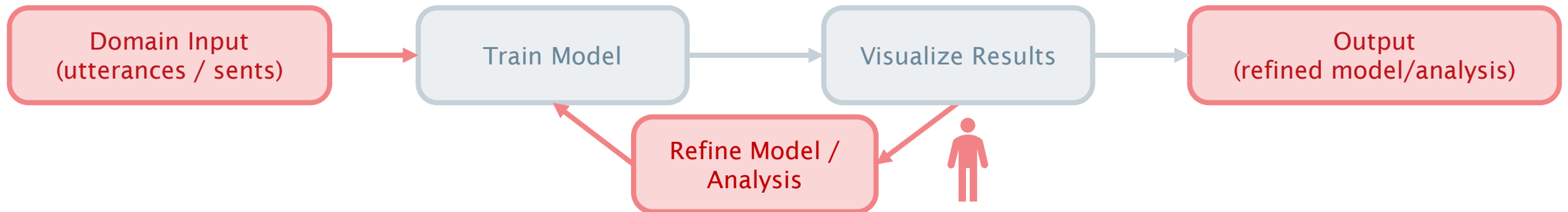
[Nie 2020] adversarial training - human generates uncertain data
[Kanchinadam 2020] active learning - model selects uncertain data

14
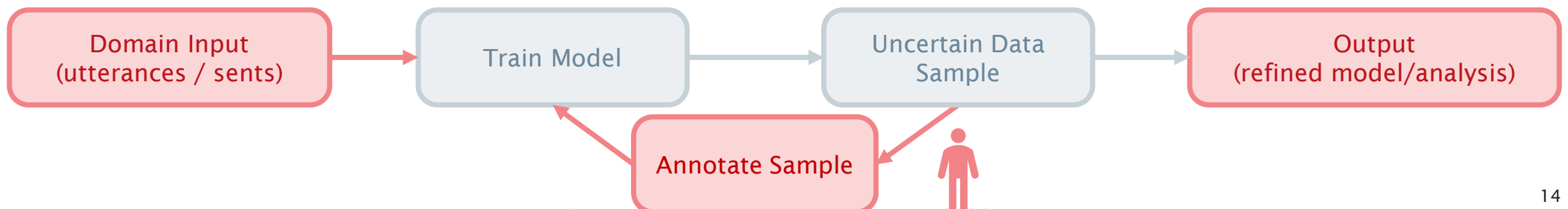
SafeSpacesNLP - Computer Science

- CLPsych 2022 shared task [May 2022]

  - https://www.aclweb.org/portal/content/clpsych-shared-task-2022

  - Task A - Post-level sequential classification of moments of change in mood

  - Task B - User-level classification of user risk based on changes in mood

  - Our models (experiments underway)

    - aggregated posts of conversation -> RoBERT embed -> MLP

    - post (within conversation) -> RoBERT embed -> sent embed -> BiLSTM

    - post (within conversation) -> RoBERT + deepemoji embed -> sent embed -> multi-task with shared BiLSTM encoder

SafeSpacesNLP - Computer Science

- SafeSpacesNLP experiments [June 2022+]

  – Baseline models from CLPsych 2022 shared task

  – New graph-augmented NLP model to encode more structure of conversations

    - Classify 'moments of change' in mental health conversations

    - Early warning for Kooth moderators to allow triage and early interventions

  – Active learning & adversarial training to explore how human-in-the-loop AI training with stakeholders can better develop in AI models

    - Southampton summer intern (active learning)

    - Nottingham summer intern (mental health trajectories within data, Kooth insitu)

    - Better models via co-training

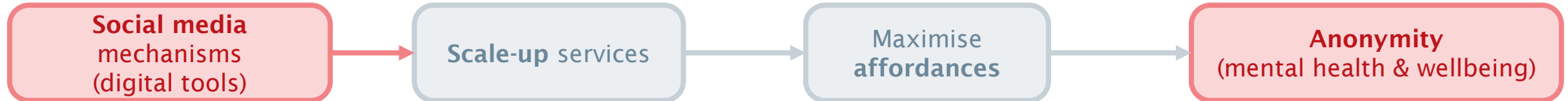   - Develop stakeholder trust in AI models over time [Middleton 2022]

Challenges (behaviour classification NLP; human in the loop NLP)

- Behaviour classification NLP

  - **Temporal context is important** for behaviour classification. Can we embed conversations into long conversation graphs? Represent fast/slow paced moments of change in behaviour?

  - Can model layer training be constrained to force models to **encode more human-understandable concept patterns** such as explanation-based training?

- Human in the loop NLP

  - Human **teams include different expertise** e.g. domain experts, AI, social science. Tailored design patterns are needed for human-in-the-loop methodologies to optimize for different contributions types e.g. AI experts could change model code on the fly if needed

  - Human-in-the-loop **evaluation metrics vary a lot**, so it can be hard to compare across approaches i.e. mixture of quantitative and qualitative

  - Often **data samples are small** with human-in-the-loop approaches. How can we identify/report bias from small sample sets? is the effect of human error magnified?

  - **Trust in AI is often difficult to build**. Can human-in-the-loop systems help stakeholders develop trust in the system we build?
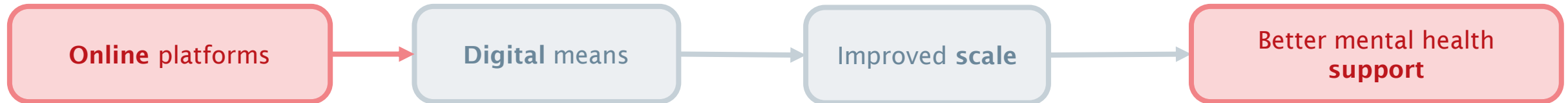
Current Research into Socio-Behaviour Analysis for Mental Health

- Socio-Behaviour Analysis for Mental Health

[McCosker, 2018]

| **Social media** mechanisms (digital tools) | → | **Scale-up** services | → | Maximise **affordances** | → | **Anonymity** (mental health & wellbeing) |

[Kivitz, 2013; National Institute of Mental Health, 2017; McCosker, 2018]

| **Online** platforms | → | **Digital** means | → | Improved **scale** | → | Better mental health **support** |

[McCosker & Darcy, 2013; Moorhead et al., 2013; Tucker and Goodings, 2017]

| Digital **technologies** | → | **Benefits** | → | Information **access**, **empowerment** & supportive **relationships** |

18

University of **Southampton**
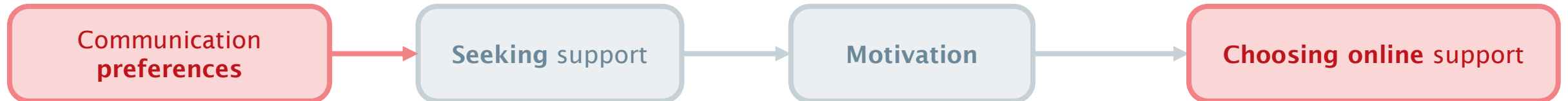
Current Research into Socio-Behaviour Analysis for Mental Health

- Socio-Behaviour Analysis for Mental Health

[Tanis, 2008; Sokol and Fisher, 2016]
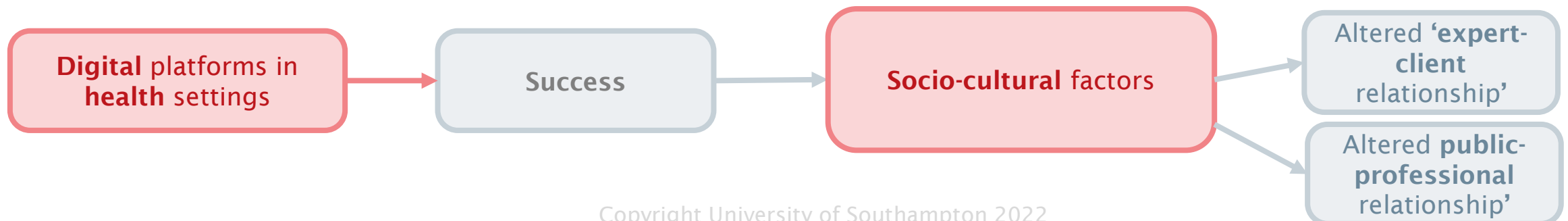
Digital **technologies** → Higher **engagement** → Maximise **reach** → Reaching the **'hard to reach'**

[Lederman et al., 2014]

Communication **preferences** → **Seeking** support → **Motivation** → **Choosing online** support
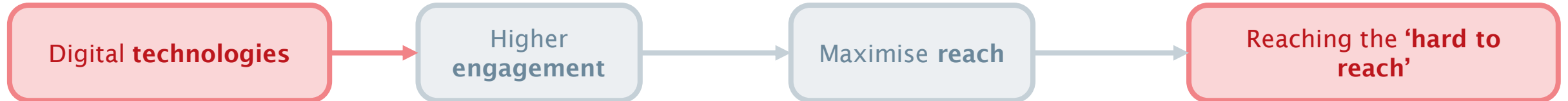
[Hansen & Aranda, 2012; Kivitz, 2013; McCosker 2018; Mowlabocus et al., 2015]

Digital platforms in **health** settings → **Success** → **Socio-cultural** factors → Altered **'expert-client** relationship'

Altered **public-professional** relationship'

19

Current Research into Socio-Behaviour Analysis for Mental Health

- Methodologies
  - Interviews, Direct observations, Questionnaires
    - e.g. structured, semi-structured
  - Content analysis
    - e.g. open, axial, selective coding, with or without software
  - Sampling tools
    - e.g. snowballing, representative
  - Structural equation modelling to test hypotheses
    - e.g. Correlation between variables; Argument quality; Verification; Literacy competence and reference credibility; Crowd consensus

SafeSpacesNLP - Applied Linguistics / Social Science

- Analysis setup - Kooth Plc
  - Free online **confidential** service, offering **counselling** and emotional wellbeing **support** to **young** people in the **UK** through **digital** platform (self-help materials, support or advice, moderated forums, personal journals, chat, drop-ins)
  - **Key actors** interviewed: 6 moderators, 3 counsellors, 3 clinical leads
  - **Practices** and **perceptions** of key digital intermediaries: main **challenges** and **risk behaviours**

  - Coding Schema motivated by a QMUL/Turing annotation schema (publication pending): "The **goal** of the Turing AI Fellowship is to create **temporally sensitive models** for capturing changes in individuals **over time** by leveraging language and other digital content they produce, which we call **user generated content** (UGC). The methods we develop consist in representing users & their behaviour over time through their UGC, **tracking changes** in these **representations** and **summarising** such changes and the **reasons** behind them."

SafeSpacesNLP - Applied Linguistics / Social Science

- Coding scheme – annotation

CODES	SUBCODES
The actor	Self-definitions; Previous/parallel experience; Background; Tasks; Challenges; Training received; Support received
The work	Specialisation; Shifts; Team; What works; What can be improved and how; Numeric indications (users to deal with/shift; submissions/quarter, etc.); Stages of the work; Rating system
The platform	What works; What can be improved; Potential issues with (semi)automatization
Risk behaviours	Physical health; Self-harm; Sexual health; Mental health; Other
Other	Emerging issues/proxy indicators of problems; COVID-19; Links to criminal behaviour/gangs;

University of
Southampton

SafeSpacesNLP - Applied Linguistics / Social Science

- Key Results
  - Roles (e.g. experience, training)
  - Risks (e.g. urgency, signposting if needed)
  - Challenges (e.g. efficiency, interpretation, location)
- Paper under development with full analysis

University of
Southampton

SafeSpacesNLP - Applied Linguistics / Social Science

- Future Directions - Forum Data

  - Data annotation (anonymised data from Kooth forum, i.e. posts and messages, moderated and unmoderated)

    1. Create vocabulary automatically from large mental health annotated Reddit corpus (500k posts, subreddit proxy for mental health label) + symptom dataset created from sources such as MedlinePlus

    2. Automatic annotation of dataset (Reddit dataset and Kooth) using vocabulary >> manual **check and correct protocol** to improve quality (CLAMP tool)

    3. Manual annotation of **moments of change** in conversation threads

    4. Manual annotation **mood** associated with post

  - Experiments planned (frequent topics/themes to pinpoint current challenges and possible solutions)

Challenges

- Digital intermediaries and 'vulnerable publics'

  - **Blurred** lines: healthcare and 'feeling management' (Hochschild, 2003)

  - Affective, emotional and **immaterial labour** (e.g., McCosker and Darcy, 2013; McCosker, 2018)

  - Facilitate **information flow**, avoid marginalisation, social and health exclusion, and stigma (Long et al., 2013), sustain online communities, help framing and reframing difficult lived experiences, and bridge users-professionals (McCosker, 2018)

  - Need to maintain **authority**, be perceived as authentic, create and maintain trust (McCosker, 2018)

  - Issues/**risks**: training (Hendry et al., 2017), working conditions, toxic content (Gillespie, 2018)

SafeSpacesNLP Summary

- SafeSpacesNLP project

  – UKRI TASHub agile project; Sept 2021 to Aug 2022
  https://www.tas.ac.uk/safespacesnlp/

  – Behaviour classification NLP in a socio-technical NLP setting

  – Online harmful behaviours for children and young people

  – Kooth >> UK's leading mental health forum provider for young people; 200,000 post dataset

- Graph-based NLP + Human-in-the-loop NLP + Socio-behaviour analysis

- Challenges

  – Behaviour classification >> Temporal context; encoding human-understandable patterns ...

  – Human-in-the-loop NLP >> Hard to compare approaches; Human error magnified ...

  – Digital intermediaries and 'vulnerable publics' >> Blurred lines in online environments user/professionals; Help reduce stigma around mental health; Risks such as toxic content ...

26

Thank you for your attention!

Dr Stuart E. Middleton[1], Dr Elena Nichele[2]

[1]University of Southampton, Electronics and Computer Science

[2]University of Nottingham, School of Computer Science

email: sem03@soton.ac.uk

        elena.nichele@nottingham.ac.uk

web: www.ecs.soton.ac.uk/people/sem03

        www.nottingham.ac.uk/computerscience/people/elena.nichele

twitter: @stuart_e_middle

References

Abdul-Mageed, M. Ungar, L. EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks, ACL 2017

Basile, V. Bosco, C. Fersini, E. Nozza, D. Patti, V. Pardo, F.M.R. Rosso, P. Sanguinetti, M. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter, SemEval-2019

Bunch, E. You, Q. Fung, G. Human-In-The-Loop Topic Discovery with Embedded Text Representations. 1st Workshop on Data Science with Human in the Loop, DaSH@KDD 2020, 2020

Caselli, T. Basile, V. Mitrović, J. Kartoziya, I. Granitzer, M. I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language, LREC 2020

Davidson, S. Sun, Q. Wojcieszak, M. Developing a New Classifier for Automated Identification of Incivility in Social Media. Workshop on Online Abuse and Harms, EMNLP 2020

Davidson, T. Warmsley, D. Macy, M. Weber, I. Automated Hate Speech Detection and the Problem of Offensive Language, ICWSM 2017

Fortuna, P. Nunes, S. A Survey on Automatic Detection of Hate Speech in Text. ACM Comput. Surv. 51, 4, Article 85, 2018

Gillespie, T. Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media. New Haven, CT: Yale University Press, 2018

Hanafi, M.F. Abouzied, A. Danilevsky, M. Li, Y. WhyFlow: Explaining Errors in Data Flows Interactively. 1st Workshop on Data Science with Human in the Loop, DaSH@KDD 2020, 2020

References

Hendry, N.A. Robards, B. Stanford, S. Beyond social media panics for 'at risk' youth in mental health practice. In: Stanford S, Sharland E and Heller NR (eds.) Beyond the Risk Paradigm in Mental Health Policy and Practice. Basingstoke: Palgrave Macmillan, pp.135-154. 2017

Hochschild, A.R. The Commercialization of Intimate Life. Berkeley, CA: University of California Press. 2003

Inches, G. Crestani, F. Overview of the International Sexual Predator Identification Competition at PAN-2012, CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, 17-20, 2012

Kanchinadam, T. Westpfahl, K. You, Q. Fung, G. Rationale-based Human-in-the-Loop via Supervised Attention, 1st Workshop on Data Science with Human in the Loop, DaSH@KDD 2020, 2020

Kayi, E.S. Diab, M. Pauselli, L. Compton, M. Coppersmith, G. Predictive Linguistic Features of Schizophrenia, SEMEVAL 2017

Khanpour, H. Caragea, C. Fine-Grained Emotion Detection in Health-Related Online Posts, EMNLP 2018

Kivitz, J. E-Health and renewed sociological approaches to health and illness. In: Orton-Johnson K and Prior N (eds.) Digital Sociology: Critical Perspectives. Basingstoke: Palgrave Macmillan, pp, 213-226. 2013

Kurrek, J. Saleem, H.M. Ruths, D. Towards a Comprehensive Taxonomy and Large-Scale Annotated Corpus for Online Slur Usage, Workshop on Online Abuse and Harms, EMNLP 2020

Long, J.C. Cunningham, F.C. Braithwaite, J. Bridges, brokers and boundary spanners in collaborative networks: a systematic review. BMC Health Services Research 13:158. 2013

McCosker, A. Engaging mental health online: Insights from beyondblue's forum influencers. New Media & Society 20(12):4748-4764. 2018

University of
Southampton

References

McCosker, A. Darcy, R. Living with cancer: affective labour, self-expression and the utility of blogs. Information, Communication & Society 16(8):1266-1285. 2013

McCosker, A. Wilken, R. Mapping mental health intermediaries: vulnerable publics and platformed support. Paper presented at AoIR 2017: the 18th annual conference of the Association of Internet Researchers, Tartu, 18-21 October. Available at: http://spir.aoir.org. 2017

Middleton, S.E. Lavorgna, L. Neumann, G. Whitehead, D. Information Extraction from the Long Tail: A Socio-Technical AI Approach for Criminology Investigations into the Online Illegal Plant Trade, WebSci '20 Companion, 2020

Middleton, S.E. Letouzé, E. Hossaini, A. Chapman, A. Trust, regulation, and human-in-the-loop AI: within the European region, Communications of the ACM (CACM), 65, 4 (April 2022), 64–68. DOI:https://doi.org/10.1145/3511597

Mowlabocus, S. Harbottle, J. Tooke, B. 'Because even the placement of a comma might be important': expertise, filtered embodiment and social capital in online sexual health promotion. Convergence 21(3):375-387. 2015

National Institute of Mental Health, Technology and the future of mental health treatment, February. Available at: https://www.nimh.nih.gov/health/topics/technology-and-the-future-of-mental-health-treatment/index.shtml. 2017

Nie, Y. Williams, A. Dinan, E. Bansal, M. Weston, J. Kiela, D. Adversarial NLI: A New Benchmark for Natural Language Understanding, ACL 2020

Park, S. Kim, D. Oh, A. Conversation Model Fine-Tuning for Classifying Client Utterances in Counseling Dialogues. NAACL-HLT 2019

Pirina, I. Çöltekin, C. Identifying Depression on Reddit: The Effect of Training Data, Social Media Mining for Health Applications Workshop & Shared Task, EMNLP 2018

References

Price, I. Gifford-Moore, J. Flemming, J. Musker, S. Roichman, M. Sylvain, G. Thain, N. Dixon, L. Sorensen, J. Six Attributes of Unhealthy Conversations, Workshop on Online Abuse and Harms, EMNLP 2020

Rashkin, H. Smith, E.M. Li, M. Boureau, Y. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. ACL 2019

Röttger, P. Vidgen, B. Nguyen, D. Waseem, Z. Margetts, H. Pierrehumbert, J. HateCheck: Functional Tests for Hate Speech Detection Models, ACL 2021

Samghabadi, N.S. Hatami, A. Shafaei, M. Kar, S. Solorio, T. Attending the Emotions to Detect Online Abusive Language, Workshop on Online Abuse and Harms, EMNLP 2020

Sharma, A. Miner, A. Atkins, D. Althoff, T. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. EMNLP 2020

Sokol R and Fisher E (2016) Peer support for the hardly reached: a systematic review. American Journal of Public Health 106(7):e1-e8.

Sosea, T. Caragea, C. CancerEmo: A Dataset for Fine-Grained Emotion Detection, EMNLP 2020

Tanis, M. Health-related on-line forums: what's the big attraction? Journal of Health Communication 13(7):698-714. 2008

Tigunova, A. Yates, A. Mirza, P. Weikum, G. CHARM: Inferring Personal Attributes from Conversations. EMNLP 2020

Tucker, I.M. Goodings, L. Digital atmospheres: affective practices of care in Elefriends. Sociology of Health & Illness 39(4):629-642. 2017

References

Van Hee, C. Jacobs, G. Emmery, C. Desmet, B. Lefever, E. Verhoeven, B. De Pauw, G. Daelemans, W. Hoste, V. Automatic detection of cyberbullying in social media text, PLoS ONE 13(10): e0203794. https://doi.org/10.1371/journal.pone.0203794

Vidgen, B. Hale, S. Guest, E. Margetts, H. Broniatowski, D. Waseem, Z. Botelho, A. Hall, M. Tromble, R. Detecting East Asian Prejudice on Social Media, Fourth Workshop on Online Abuse and Harms, EMNLP 2020

Waseem, Z. Hovy, D. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter, NAACL Student Research Workshop, NAACL 2016

Wulczyn, E. Thain, N. Dixon, L. Ex Machina: Personal Attacks Seen at Scale. WWW 2017

Xu, J. Szlam, A. Weston, J. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. arXiv 2021

Zirikly, A. Resnik, P. Uzuner, O. Hollingshead, K. CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts, NAACL 2019