

**PRÉDICTION DE LA DÉTÉRIORATION DU
COMPORTEMENT À L'AIDE DE L'APPRENTISSAGE
AUTOMATIQUE**

par

Jean Marie Kabamba Tshimula

Thèse présentée au Département d'informatique
en vue de l'obtention du grade de philosophiæ doctor (Ph.D.)

FACULTÉ DES SCIENCES
UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, 17 août 2022

Le 17 août 2022

*Le jury a accepté la thèse de Jean Marie Kabamba Tshimula dans sa
version finale*

Membres du jury

Professeur Shengrui Wang

Directeur de recherche

Département d'informatique

Professeur Belkacem Chikhaoui

Codirecteur de recherche

Département Science et Technologie

Université TÉLUQ

Richard Khoury

Professeur

Membre externe

Département d'informatique et de génie logiciel

Université Laval

Professeur Manuel Lafond

Membre interne

Département d'informatique

Professeur Marc Frappier

Président-rapporteur

Département d'informatique

Sommaire

Les plateformes de médias sociaux rassemblent des individus pour interagir de manière amicale et civilisée tout en ayant des convictions et des croyances diversifiées. Certaines personnes adoptent des comportements répréhensibles qui nuisent à la sérénité et affectent négativement l'équanimité des autres utilisateurs. Certains cas de mauvaise conduite peuvent initialement avoir de petits effets statistiques, mais leur accumulation persistante pourrait entraîner des conséquences majeures et dévastatrices. L'accumulation persistante des mauvais comportements peut être un prédicteur valide des facteurs de risque de détérioration du comportement. Le problème de la détérioration du comportement (DC) n'a pas été largement étudié dans le contexte des médias sociaux.

La détection précoce de la DC peut être d'une importance cruciale pour éviter que le mauvais comportement des individus ne s'aggrave. Cette thèse aborde le problème de la DC dans le contexte des médias sociaux. Nous proposons de nouvelles méthodes basées sur l'apprentissage automatique qui (1) explorent les séquences comportementales et leurs motifs temporels pour faciliter la compréhension des comportements manifestés par les individus et (2) prédisent la DC à partir de combinaisons consécutives de motifs séquentiels correspondant à des comportements inappropriés. Nous menons des expériences approfondies à l'aide d'ensembles de données du monde réel et démontrons la capacité de nos modèles à prédire la DC avec un haut degré de précision, c'est-à-dire des scores F-1 supérieurs à 0,8. En outre, nous examinons la trajectoire de DC afin de découvrir les états émotionnels que les individus présentent progressivement et d'évaluer si ces états émotionnels conduisent à la DC au fil du temps. Nos résultats suggèrent que la colère pourrait être un état émotionnel potentiel qui pourrait contribuer substantiellement à la détérioration du comportement.

Mots-clés: Affinité, Séquences comportementales, États émotionnels, Caractéristiques psycholinguistiques, Personnalité, Détérioration, Apprentissage automatique.

Acknowledgements

I would like to start my acknowledgements by thanking my advisors, Dr. Shengrui Wang and Dr. Belkacem Chikhaoui, for intellectual discussions, and for helping me to navigate the arcane of academia. Both of them have helped me selflessly and provided me so many inspirations, much advice, and endless support.

Thanks to Dr. Marc Frappier and Dr. Manuel Lafond, my viva examiners, for their close reading of this dissertation. After 4 years of work, I appreciated the time you took to give me a rigorous and engaging examination.

I would like to express my warmest gratitude to those who have accompanied and shared their knowledge with me at the ProspectUS Laboratory: Philippe Chatigny, Heng Shi, Rongbo Chen, Patrick Owusu, Kunpeng Xu, Mingxuan Sun, Abdallah Aaraba, Théodore Simon, Olfa Gassara, Imen Montassar, and many others.

I would like to thank Mme Carol Harris for linguistic assistance on all my research publications.

Thanks to Dr. Jianfei Zhang, Dr. Annie Carrier, Dr. Denis Bédard, Dr. Sharmistha Gray, Dr. Esaie Kuitche, Dr. Étienne Tajeuna, Dr. Jean Paul Faye, Gwen Andrews, Achraf Essemlali, Didier Mbuyi, DJeff Kanda, Hugues Kanda, Patrick Lutumba and Patrick Bungama for having time for discussions on work and asking me thought-provoking questions.

I would like to thank Dr. Marine Hadengue, Philippe Arbour and the Arbour Foundation for granting me a scholarship.

Thanks to my parents for all your love and support in making the right choices, and for being proud of what I do.

I am deeply indebted to my wife Naomie. She has endured much because of me. She has devoted herself to me and our children, Jonel-Alvin and Helena-Joyce. I dedicate this milestone to you, the love of my life.

Table of Contents

Sommaire	iii
Acknowledgements	iv
Table of Contents	v
List of Figures	viii
List of Tables	x
List of Abbreviations	xii
Introduction	1
1 Context and Problems of Behavioral Deterioration	4
1.1 Motivation	4
1.2 Research challenges	8
1.3 Research contributions	10
I Understanding Emotional States using Psycholinguistic Features	17
2 Investigating Moral Foundations from Web Trending Topics	20
2.1 Introduction	21
2.2 Related work	23
2.3 Methodology and data processing	24

TABLE OF CONTENTS

2.4	Discussion	26
2.5	Conclusion	28
3	Emotion Detection in Law Enforcement Interviews	29
3.1	Introduction	30
3.2	Related work	31
3.3	Methods	32
3.4	Experiment	34
3.5	Results	37
3.6	Ethical considerations	43
3.7	Discussion and conclusion	44
II	Discovery of Affinity and Personality from Text Data	46
4	A New Approach for Affinity Relationship Discovery in Online Forums	49
4.1	Introduction	50
4.2	Related work	52
4.3	Proposed method	54
4.4	Data preparation	60
4.5	Experiments	62
4.6	Conclusion and future work	73
5	Discovering Affinity Relationships between Personality Types	74
5.1	Introduction	75
5.2	Related work	77
5.3	Datasets	79
5.4	Methodology	80
5.5	Experiments	83
5.6	Discussion	90
5.7	Conclusion	93

TABLE OF CONTENTS

III Stance Detection and Behavioral Deterioration in Discussions	95
6 A Pre-training Approach for Stance Classification in Online Forums	98
6.1 Introduction	99
6.2 Related work	101
6.3 Model	102
6.4 Experiments	105
6.5 Results and discussion	109
6.6 Psychological processes	112
6.7 Conclusion	113
7 On Predicting Behavioral Deterioration in Online Discussion Forums	115
7.1 Introduction	116
7.2 Related work	118
7.3 Model	121
7.4 Experimental setup	123
7.5 Results and discussion	124
7.6 Conclusion	127
8 Discovery of Temporal Deterioration Patterns from Behavioral Sequences	129
8.1 Introduction	129
8.2 Method	132
8.3 Deterioration prediction	139
8.4 Emotional states and deterioration	143
8.5 Discussion and conclusion	146
Conclusion	148
Bibliography	179

List of Figures

1.1	An overview of all problems studied in this dissertation.	6
1.2	Trajectory of behavioral deterioration	9
3.1	Top 30 tri-grams occurring in BGC and GHC.	38
3.2	Top 30 tri-grams occurring in LRC and RWC.	39
3.3	Temporal evolution of emotional states during interviews.	40
3.4	Similarity between the emotional trajectory of the suspect and detectives.	42
4.1	HAR graph (positive interaction sequence-based transition diagram).	63
4.2	Distribution of affinity score ranges by experimental datasets.	64
4.3	Evolution of affinity relationships and social groups over time.	67
5.1	Affinity percentages between the 136 combinations of the 16 MBTI personality types.	84
7.1	Feature extraction process used to capture deterioration patterns within behavioral sequences (BS).	122
7.2	Results of behavioral deterioration prediction by adding extra features to the main model.	125
8.1	Illustration of a behavioral sequence	132
8.2	Illustration of the alignment of community behavioral trajectories in a behavioral matrix	134
8.3	Overview of the bidirectional LSTM model	137
8.4	Architecture of BiLSTM with attention	138

LIST OF FIGURES

8.5	Results of the individual level prediction of behavioral deterioration by adding extra features to the main model.	141
8.6	Results of the community level prediction of behavioral deterioration by adding extra features to the main model.	142

List of Tables

1.1	Some types of misbehavior.	5
2.1	Words utilized for prediction. We arbitrarily picked 15 words per dimension from the MFD-MoralStrength dictionary.	23
2.2	Pearson correlation (r) between moral word scores and LIWC features during coronavirus lockdown in Canada.	25
2.3	Pearson correlation (r) between moral word scores and LIWC features during a specific period of WEXIT.	25
2.4	Classification results for moral foundations using the combination of ZSC and predictive models.	27
4.1	An example to illustrate the functioning of HAR-search.	57
4.2	Dataset Information.	61
4.3	Performance results for the proposed method and baselines on four experimental datasets.	71
5.1	Data summary and distribution.	79
5.2	Semantic similarity for affinity relationships between different MBTI personality types.	85
5.3	Pearson correlations between LIWC (positive emotions) features extracted on language use to discover emotional stability in affinities between two different personality types.	86
5.4	Pearson correlations between LIWC (negative emotions) features extracted on language use to discover emotional stability in affinities between two different personality types.	87

LIST OF TABLES

5.5	Prediction results of MBTI personality types.	88
5.6	Clustering results in terms of Error and NMI.	89
5.7	Characteristics of the MBTI types.	94
6.1	An example to demonstrate the functioning of our approach.	103
6.2	An example to illustrate feature extraction.	104
6.3	Stance classification results for the proposed method and baselines on the two experimental datasets.	109
6.4	Quantifying divergence of opinion by topic.	110
6.5	Prediction performance (Pearson’s r) based on 10-fold cross-validation using LIWC features (positive and negative emotions) extracted from different topics addressed on IAC2 and ACD datasets.	112
7.1	Results of behavioral deterioration prediction.	124
8.1	Results of the individual level prediction of behavioral deterioration. .	139
8.2	Results of the community level prediction of behavioral deterioration.	140
8.3	Prediction quality for emotional states at the individual level, as measured using the Pearson r . The results concern sub-datasets indicating the presence of deterioration patterns.	143
8.4	Prediction quality for emotional states at the community level, as measured using the Pearson r . The results concern sub-datasets indicating the presence of deterioration patterns.	143
8.5	Prediction quality for emotional states at the individual level, as measured using the Pearson r . The results concern sub-datasets indicating the absence of deterioration patterns	144
8.6	Prediction quality for emotional states at the community level, as measured using the Pearson r . The results concern sub-datasets indicating the absence of deterioration patterns	144

List of Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
BS	Behavioral Sequences
DTW	Dynamic Time Warping
HAR	Hidden Affinity Relationships
LEI	Law Enforcement Interviews
LDA	Latent Dirichlet Allocation
LIWC	Linguistic Inquiry and Word Count
LR	Logistic Regression
LSTM	Long Short-Term Memory
MBTI	Myers-Briggs Type Indicator
MCL	Markov Cluster Algorithm
MFD	Moral Foundations Dictionary
MRC	Medical Research Council
NLI	Natural Language Inference
NLP	Natural Language Processing
NRC	National Research Council of Canada
OSN	Online Social Network
PIS	Positive Interaction Sequences
RoBERTa	A Robustly Optimized BERT Pre-training Approach
RF	Random Forest
SVM	Support Vector Machine
ZSC	Zero-Shot Text Classification

Introduction

Social media platforms assemble individuals who have diversified convictions and beliefs to interact in friendly and civilized ways, most of the time. Increasingly, however, they are having the opposite behavior, due to a rising tide of deviations, and deliberate provocations; since some individuals engage in misbehavior that undermines serenity and adversely affects the equanimity of other users. Some instances of misbehavior may initially have small statistical effects, but their persistent accumulation may subsequently have major and devastating consequences. For example, some victims of cyberbullying are more likely to self-harm, engage in suicidal behavior, and experience some unpleasant aftermaths, including psychological and anxiety disorders; others even commit suicide. The persistent accumulation of misbehavior can be a valid predictor of risk factors for behavioral deterioration. The problem of behavioral deterioration has not been widely studied in the context of social media. Early detection of deteriorating behavior is critically important to prevent individuals' misbehavior from escalating in severity.

This dissertation aims to develop machine learning models to address the above-mentioned problem. Consequently, we divided this problem into three components, namely affinity, personality, and deterioration, to investigate the underlying factors contributing to a behavioral deterioration to preemptively detect and predict behavioral deterioration. More specifically, we propose the following techniques: (1) a new approach based on natural language inference that utilizes psycholinguistic features to discover whether individuals who commit misbehavior exhibit social morality and emotional instability. (2) an advanced method based on Markov models, machine learning, and natural language processing to quantify affinity scores; investigate affinity over time, and predict affinity relationships arising from the influence of certain users. (3) a new approach based on machine learning to better identify the influence

of personality on affinity. The model identifies among the affinity relationships the personality types that seem to foment misbehavior within social media platforms. (4) a novel approach based on machine learning to construct behavioral sequences (BS) from the set of temporal behaviors exhibited by individuals and predict behavioral deterioration at the individual and community level from consecutive combinations of sequential patterns within BS. Additionally, we investigate the trajectory of behavioral deterioration within BS to discover the emotional states that individuals manifest and to assess whether these emotional states contribute to behavioral deterioration over time.

We demonstrate the effectiveness of the proposed models through several scopes on real-world datasets and from different horizons. Our results indicate that our models have the potential to leverage behavioral sequences to predict behavioral deterioration at the individual and community level and show the ability of our models to predict behavioral deterioration with a high degree of accuracy, i.e., F-1 scores of over 0.8. Our findings suggest that *anger* could be a potential emotional state that can substantially contribute to behavioral deterioration.

This dissertation is structured as follows. Apart from the conclusion and Chapter 1, we organize the rest of chapters in Parts I, II and III:

- Chapter 1 highlights the importance of investigating the problem of behavioral deterioration and proposes three components to address this problem.
- Part I deals with understanding social morality and emotional states from text data using psycholinguistic features. Chapter 2 examines moral foundations from large-scale social media text data from trending topics, predict moral values and investigate whether differences in moral values have a certain influence on emotional traits. Chapter 3 investigates the temporal evolution of emotional states and identifies relevant patterns that are relevant to an emotional breakdown.
- Part II discovers affinity and personality using text data. Chapter 4 addresses the problem of discovering affinity relationships and predicts the evolution of affinity and affinity relationships arising from the influence of certain users. Chapter 5 investigates the influence of personality types on affinity, measures emotional stability and semantic similarity between affinity relationships and

INTRODUCTION

then predicts personality from text data.

- Part [III](#) addresses stance detection and behavioral deterioration in discussions. Chapter [6](#) explores how the divergence of opinion can potentially conduct unhealthy conversations and emotional reactions. Chapters [7](#) and [8](#) introduce a formal definition of the problem of behavioral deterioration and predict behavioral deterioration at the individual and community level. Chapter [8](#) evaluates the emotional states that contribute to behavioral deterioration.
- The conclusion summarizes the contributions addressed in this dissertation and different application perspectives and discusses several potential directions for future work.

Chapter 1

Context and Problems of Behavioral Deterioration

In this chapter we highlight the importance of investigating the problem of behavioral deterioration. We begin with the motivation and research challenges, followed by problem setup and details of our solution and its scope. In the end we report a series of contributions addressed in this dissertation.

1.1 Motivation

Social media has become an important resource for investigating user behaviors through their digital footprints as it provides a popular space in which numerous topics are discussed between people who may be like-minded or hold opposing views. In social media, some people show common sense, tolerance, and respect for the views of the online community members, while others manifest intransigent attitudes and engage in misbehavior that harms the community and adversely affects the equanimity of online community members. Misbehavior includes but is not limited to abusive and offensive language, threats, hate speech, cyberbullying, and race and gender discrimination (Table 1.1) and can be expressed through a post in various ways, such as texts, videos, pictures, taunting emoticons, etc. Misbehavior may refer to inappropriate behavior, disruptive behavior, and/or maladaptive behavior characterized by covert or overt hostility and intentional aggression towards others. Some instances

1.1. Motivation

Table 1.1 – Some types of misbehavior.

Type	Reference
– Cyberbullying	[75, 154, 159]
– Hate speech <ul style="list-style-type: none"> — Racism, homophobia, or other behavior that discriminates against a group or class of people — Genre discrimination — Sexual harassment of any kind, such as unwelcome sexual advances or words/actions of a sexual nature 	[41, 106, 124]
– Offensive or abusive language <ul style="list-style-type: none"> — Abusive action directed at an individual, such as threats, intimidation, or bullying 	[41, 197, 203, 216]
– Trolls, unfair generalization and sarcasm	[119, 151]

of misbehavior may initially have small statistical effects, but their persistent accumulation may subsequently have major and devastating consequences. For example, some victims of cyberbullying are more likely to self-harm, engage in suicidal behavior [50, 94], and experience some unpleasant aftermaths, including psychological and anxiety disorders [40, 87, 120]; others even commit suicide [79].

Misbehavior can escalate to violent behavior when the perpetrators constantly harm others and do not get sanctioned for their misdeeds. However, sanctioning moral transgression and norm violations may be an important aspect to shape and keep healthy the online discussion community against deviations and deliberate provocations [151]. Violent behavior may consequently be considered as the endpoint on a continuum of behavioral deterioration [52].

CHAPTER 1. CONTEXT AND PROBLEMS OF BEHAVIORAL DETERIORATION

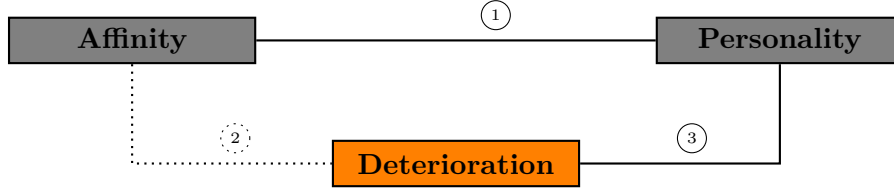


Figure 1.1 – An overview of all problems studied in this dissertation.

Behavioral deterioration has not been extensively studied in the context of online discussion communities. Deterioration may be defined in different ways, and regardless of the definition, it is difficult to measure. Specifically, we define deterioration as the accumulation of misbehaviors. Early detection of behavioral deterioration can be of crucial importance in preventing individuals’ misbehavior from escalating in severity.

In order to investigate the trajectory underlying behavioral deterioration (see Figure 1.1), we divide the problem into three specific components as follows:

① **Affinity–Personality discovery.** We investigate whether affinity relationships are associated with certain types of personality. Basically, the degree of affinity can be perceived as a score that indicates the close proximity between individuals in the relationship they share. Drawing the connection between affinity and personality may help discover the type of individuals that one prefers to be friends with, and portray what one is like. It may be conscious or unconscious that one always develops affinity relationships with individuals who fall into the same personality traits as them. We treat this possibility from different angles to consolidate observed patterns from language usage, idiosyncrasy and psychological traits to discover whether it is purely coincidental or a simple means that an individual utilizes to maintain her comfort zone. As a result, we build models that can predict personality from individual postings, involving multiple interacting factors that may contribute to personality, such as mental state, affinity and capacity to deal with divergent opinions expressed on the topics addressed within the community. Our models can identify among the affinity relationships those individuals who seem to foment misbehavior within the community; and assess the likelihood that their affinity may evolve over time and the risks they

1.1. Motivation

may represent.

② **Deterioration.** We introduce a formal definition of the problem of behavioral deterioration and predict signals relevant to deterioration from consecutive accumulations of behaviors exhibited by individuals within a discussion forum.

③ **Personality–Deterioration discovery.** A person’s behavior may undergo a sudden or gradual deterioration, and this phenomenon may happen under the influence of friends, self-commitment, reckoning, or because of some topics addressed or some mental health conditions. We examine the trajectory of behavioral deterioration in order to discover emotional states that individuals gradually display and evaluate whether it contributes towards dramatically worsening the behavior over time. However, personality changes that are uncontrollable, uncomfortable, and detrimental may be a sign of a deeper problem. Owing to this, we scrutinize emotional states that may affect personality change, including *anxiety*, *stress*, *fear*, *anger*, *sadness*, *disgust*, and *surprise*. Specifically, we analyze emotional states that have a greater relationship with deterioration; and investigate their effects and impulsiveness that they engender on the occurrence of behavioral deterioration.

This dissertation contributes to the production of literature on behavioral deterioration and opens the door for promising directions of future research. Our models for building behavioral sequences and predicting behavioral deterioration from consecutive combinations of sequential patterns are practically useful to a variety of domains such as clinical psychology, computational science, social science and education. Moreover, our discoveries can be used by companies, schools, prisons, psychiatric centers, and organizations for monitoring people manifesting signals relevant to behavioral deterioration; for instance, psychiatric centers can utilize our models to track the consecutive accumulation of daily signs of individuals with mental health conditions to predict signals relevant to deterioration or improvement. In prisons, our models can be used to predict the behavioral deterioration of recidivists and inmates stimulating defiant and aggressive behaviors. At schools, our models can be utilized as a barometer to measure behavior escalation and predict negative affinity relationships and behavioral deterioration from breaking the behavior code and student misbehaviors such as disruptive talking, chronic avoidance of work, clowning,

CHAPTER 1. CONTEXT AND PROBLEMS OF BEHAVIORAL DETERIORATION

interfering with teaching activities, harassing classmates, verbal insults, rudeness to teacher, defiance, hostility, absenteeism, bullying and other inappropriate behaviors. In companies, our models can be applied to predict the behavioral deterioration of employees engaged in code-of-conduct violations such as discrimination, gossiping, bad jokes, physical threats, negative remarks, and so on.

1.2 Research challenges

While investigating these three components to study behavioral deterioration is important, there are significant research gaps toward modeling and solving them efficiently and effectively. Here, we identify several main research challenges:

1. **Quantifying affinity relations from online discussions.** The concept of affinity relationship discovery is relatively new in the context of online discussion communities and there has been little work addressing it to date. This problem entails finding affinity relationships in a community by combining structural features, temporal information, and the content of interactions without necessarily taking into consideration offline inputs, such as the social, cultural, and psychological environment and socioeconomic status; or even social ties that users have offline. Affinity discovery seeks not only to identify these affinity relationships but also to quantify them so that the degree of affinity between individuals can be perceived in the form of a score.
2. **Annotating data for detecting deterioration.** Detecting deterioration leverages large-scale text data. Identifying subtle indicators of behavioral deterioration is a difficult task. One of the key ingredients to progress on this task is high-quality, large, and annotated datasets. One of the challenges that we face is processing data and understanding the context of the data being analyzed. Since the problem of behavioral deterioration is new in the context of social media, we also face the absence of adequate annotated data on deterioration. Alternatively, we track accumulations of different behavior classes from real-world data to investigate the trajectory of deterioration (see Figure 1.2).
3. **Defining deterioration.** Detecting behavioral deterioration on social media

1.2. Research challenges

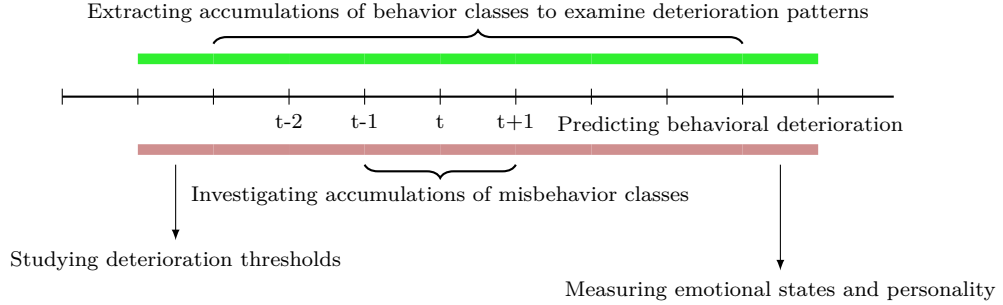


Figure 1.2 – Trajectory of behavioral deterioration. Note that each time point in the trajectory represents a specific behavior class

has not been fully explored so signals relevant to deterioration are not clearly formulated. As mentioned above, we define deterioration as the accumulation of misbehaviors. First, we build behavioral sequences from temporal behaviors exhibited by individuals within the community. We thereafter extract potential features to examine deterioration patterns within behavioral sequences. Intuitively, the model (i.e., the number of features) to predict behavioral deterioration depends heavily on the average length of the set of behavioral sequences. However, ground truth information to capture the prediction accuracy of behavioral deterioration is scarce and challenging since this problem has not been addressed in the context of social media. The lack of ground truth information does not affect the generalization of the findings and model performance, since the results stem directly from observed accumulations of behaviors exhibited by individuals within the community. Our feature extraction strategy consists of closely watching over the consecutive accumulation of misbehaviors in behavioral sequences to discover deterioration-relevant signals and investigating the rate at which these signals evolve as time moves forward. This makes intuitive sense, since we do not rely on a single label as a global context for the whole behavioral sequence, we instead regard local contexts using our feature extraction strategy to investigate the trajectory of deterioration-relevant signals. We then examine the local contexts at the global level (i.e., in the behavioral sequence) to discover the shifts of deterioration signals over time and discover the pace at which they evolve in order to predict behavioral deterioration.

CHAPTER 1. CONTEXT AND PROBLEMS OF BEHAVIORAL DETERIORATION

4. **Fixing deterioration thresholds.** Beyond monitoring accumulations of behavior classes to extract feature sets, we face challenges in defining threshold values to determine whether a set of behavioral sequences for individuals tends to deteriorate or not. Such scores could allow establishing different degrees of deterioration in order to facilitate more effective monitoring of the trajectory of behavioral deterioration.
5. **Building holistic models for measuring deterioration.** One of the challenges is to construct models that study deterioration as a whole by taking into consideration supplementary factors, including behaviors exhibited within the community. To build such models, we have to *(i)* examine correlations between language use of individuals for which behavior sequences comprise accumulations of behavior classes that indicate signals relevant to deterioration; *(ii)* analyze personality traits to understand whether deterioration occurs under the effects of the topics addressed in the discussion forum, mental health conditions or some other factors and *(iii)* understand the impact of some personal concerns (such as work, money, religion, death, etc.) on behavioral deterioration.

1.3 Research contributions

Keeping the research challenges in mind, in this dissertation we develop novel computational models for discovering, formulating, modeling, and solving the problems introduced in §1.1. Here, we describe our contributions toward affinity, personality, and deterioration. The contributions in §1.3.1 are briefly detailed in Part I (Chapters 2 and 3); §1.3.2 and §1.3.3 in Part II, Chapter 4 and Chapter 5 respectively; and §1.3.4 in Part III. Publication details are included in the summary sections of Parts I, II and III. It is important to mention that we closely work with a psychologist for the validation of results.

1.3.1 Detecting emotional states in text data

We investigate social morality to understand moral differences in a broad spectrum of interactions on social media. Morality guides human social interactions and

1.3. Research contributions

can potentially conduct to a divergence of opinion, polarity, and hostility when there is moral shock within a community. The key insight is to discover whether differences in moral dimensions (*care/harm*, *fairness/cheating*, *loyalty/betrayal*, *authority/subversion* and *purity/degradation*) have an influence on emotional states. Additionally, we examine emotional states to discern the factors that lead to emotional instability in highly motivated high-conflict interactions such as police interrogations.

Concretely, we address the two problems as follows. First, we utilize the Moral Foundations Dictionary (MFD)¹ and propose a model based on natural language inference (NLI) to automatically extract morality features that we then use for prediction using Support Vector Machine and Logistic Regression. The MFD is one of the most established dictionary methods for language analysis in psychological science and provides information on the proportions of virtue and vice words for each moral dimension. We extract psycholinguistic features in text data to measure emotional states. We compute Pearson correlations between the MFD word scores and emotional states to discover the influence of morality on emotional states. More specifically, we used the following emotional states: *positive emotion*, *negative emotion*, *anger*, *anxiety*, and *sadness*. We find that the lack of annotated data does not affect the generalizability of the findings and model performance. Our results provide strong evidence that we can predict moral foundations with an accuracy exceeding 0.65 and identify statistical significance that indicates the influence of morality over emotional states within text data. Next, we utilize psycholinguistic features to investigate the temporal evolution of emotional states and identify patterns that are relevant to an emotional breakdown. More specifically, we apply the Linguistic Inquiry and Word Count (LIWC) dictionary [144] to extract the features of these emotional states (*pessimism*, *fear*, *anger* and *optimism*), and propose a model based on NLI to quantify emotional states in order to construct an emotional trajectory for individuals participating in interactions. LIWC is a widely used psychometrically validated system for psychology-related analysis of language and word classification. We apply the Dynamic Time Warping (DTW) algorithm [88] to measure the similarity between the emotional trajectories and to identify patterns relevant to an emotional breakdown. DTW is an algorithm for measuring similarity between two temporal sequences. The

1. <https://moralfoundations.org>

closer the value is to 0, the more similar the two temporal sequences are.

Through extensive experiments on four different datasets, our findings indicate emotional trajectories illustrating shifts in emotional states and show similarities and correlations between the emotional trajectories in efficient ways.

1.3.2 Discovering affinity in online forums

The problem of affinity relationships in the context of social media has not been clearly and formally defined in the literature. To clarify this concept, we first define affinity relationships as being relationships that include a set of characteristics such as mutual understanding, reciprocal and common interests, sympathy, harmonious communication or agreement between individuals. Affinity goes beyond the conventional conception of friendship in social media (e.g., two individuals who mutually follow back on Twitter-like platforms). An affinity relationship can be detected in interactions or by the way that individuals exchange. Individuals progressively develop affinity and get closer as they mention each other frequently in interactions and share information with one another. Fundamentally, affinity can be positive or negative to some extent –for instance, if someone is always taking an opposite viewpoint from someone else in the discussion. Negative affinities may be built by sentences expressed respectfully to contradict someone without insults, etc. Such interactions are taken as positive based on the context in which they occur. However, the content of interactions contains important characteristics that can be utilized to discover potential signals relevant to affinity in the form of a score.

Concretely, we investigate the combination of structural features, temporal information, and the content of the interactions for quantifying the degree of affinity between individuals in online forums. We develop an advanced method based on Markov models, machine learning, and natural language processing to quantify affinity scores. We utilize the quantified affinity scores to track the evolution of affinity over time and predict affinity relationships arising from the influence of certain users. Specifically, we propose mathematical definitions of affinity influence and extract several interpretable features from these definitions; evidence has shown the benefits of including these features in affinity prediction.

1.3. Research contributions

Through extensive experiments on a variety of datasets, we evaluate our approach’s versatility and effectiveness in predicting affinity. We test our approach on three different classification algorithms (Support Vector Machine, Random Forest, and Logistic Regression), and in most cases, it performs well in terms of precision, recall, and F-measure. Note that we compute precision and recall measures in the absence of ground truth [109]. Our approach results in robust discovery by considering minute details and predicting affinity influence with higher accuracy and outperforms the baselines in all prediction horizons by a considerable margin. Our results show strong evidence that combining structural and temporal features and the content of interactions provides better performance on the task of affinity prediction.

1.3.3 Discovering relations between affinity and personality

We study the relationships between affinity and personality and seek to understand how individuals with similar personality traits get to develop their affinity and discern what attracts an individual to another. Psychology research findings suggest that personality is related to differences in friendship characteristics and that some personality traits correlate with linguistic behavior. Most studies in psychology conduct survey questionnaires (and/or written essays) to assess personal behavioral preferences. In this dissertation, we track language use and interactions between individuals on social media; we believe that spontaneous language contains genuine behaviors, personality traits and feelings of individuals. To efficiently conduct an analysis of language use between individuals based on their personality types, we collected a publicly available dataset containing information on individuals who self-identified with a Myers-Briggs personality type (MBTI). Specifically, the MBTI assessment is based on research and personalized preferences and can contribute important information to the understanding of individual psychological functions such as intuition, sensation, thinking, feeling, etc. The MBTI model defines four binary dimensions: (1) Introversion-Extraversion (I-E), (2) Intuition-Sensing (NS), (3) Feeling-Thinking (F-T), and (4) Perception-Judgment (P-J) that combine to yield 16 personality types into which individuals may be classified (e.g., INFP, ESTJ, ISFJ, etc). The 16 MBTI personality types are simple to manipulate to account for personality differences.

CHAPTER 1. CONTEXT AND PROBLEMS OF BEHAVIORAL DETERIORATION

Concretely, we measure emotional stability and semantic similarity between affinity pairings. To this end, we utilize the GloVe word embedding [145], an unsupervised learning algorithm for obtaining vector representations for words pertaining to specific personality types, and then apply cosine similarity to measure semantic similarity between personality types. The motivation for using GloVe is that it does not solely rely on local context information of words and focuses on words co-occurrences over the whole corpus, and its embeddings relate to the probabilities that two words appear together. We extract psycholinguistic features using the LIWC dictionary to calculate Pearson correlations related to emotional stability (i.e., *positive*, and *negative emotions*) between affinity pairings. Additionally, we utilize self-identified MBTI personality types as annotations and train five different models (Logistic Regression, Random Forest, Support Vector Machine, Naive Bayes and BERT [43]) to predict personality in the linguistic level of interactions. To detect the influence of personality on affinity, we utilize two different graph clustering techniques: the random walk hitting time-based digraph clustering algorithm (K-destinations) [28] and the Markov cluster algorithm (MCL) [191]. The motivation for using these algorithms in an affinity context is that they are based on first-order Markov chains, deal with directed graphs, and draw on intuition from random walks on graphs to detect cluster structure. Our affinity approach utilizes Markov Chain models to generate a transition matrix that quantifies affinity scores. Specifically, we apply graph clustering to discover the connectivity between nodes within each cluster and then heuristically investigate the influence of personality on affinity based on cluster results and distributions.

Through extensive experiments on a dataset of 25,253,604 tweets, our results provide some of the first insights into the investigation and understanding of affinity relationships between personality types on social media and identify several statistically significant correlations in terms of emotional stability in personality-based affinity pairings. Our results also identify influential personality types that weigh more heavily on affinity relationships. Our findings suggest that semantic similarity and emotional stability constitute an important lead for understanding the implications of personality in the development of affinity. We show that MBTI personality types can be predicted from the spontaneous language with an F-1 score superior to 0.76, a resolution that is likely fine-grained enough for various experimental datasets.

1.3. Research contributions

1.3.4 Predicting behavioral deterioration

Detecting signals relevant to deterioration can help prevent misbehavior from escalating in severity. Behavioral deterioration is the stage that individuals could reach, in perpetrating misbehavior. Behavioral deterioration can engender negative and devastating consequences for victims, including self-harm, psychological and anxiety disorders, and suicidal behavior. The problem of behavioral deterioration has not been extensively studied in the context of social media. In this dissertation, we investigate the trajectory of behavioral deterioration and propose new models that construct behavioral sequences from temporal behaviors exhibited by individuals. Before we address this problem, we analyze stances expressed by individuals in social media discussions and study how the divergence of opinion in the discussed topics (such as *Evolution*, *Gay Marriage*, *Abortion*, *Gun Control*, and *Death Penalty*) can potentially lead to unhealthy conversations and emotional reactions.

Concretely, we propose models for stance detection and behavioral deterioration prediction. For stance detection, we propose a pre-trained model of textual entailment on top of RoBERTa [122], a Transformer-based model. This model classifies stances by capturing the context of the discussion through the examination of pairs of stances and relational structures of discussion specific to each topic within the defined window of interactions of each participant of the discussion. We examine the degree of disagreement and neutrality to measure the divergence of opinion on topics addressed in the discussion. Specifically, we construct topics-based graphs to measure divergence of opinion throughout the discussion and measure the emotional states manifested in different discussion topics. To quantify divergent opinions, we used four measures of probability divergence: Kullback-Leibler [167], Jensen-Shannon [22], Hellinger distance [165], and Bhattacharyya distance [99]. To this end, we take each of these interaction pairs, apply the word2vec skip-gram model [129] to embed the words of each interaction of the pair as vectors in a low-dimensional space, and finally encode them as probability densities. One of the advantages of using skip-gram model is that it predicts related context words of a given target word and considers the order of surrounding words during training. The densities represent the distributions over the possible significations of a word. We compute the divergence metrics between the distributions of interaction pairs. To estimate divergence of opinion on

CHAPTER 1. CONTEXT AND PROBLEMS OF BEHAVIORAL DETERIORATION

the whole topic, we compute the arithmetic mean of the results obtained from all interaction pairs of interest. We also extract psycholinguistic features using to quantify emotional states in interactions. To predict the emotion associated with interactions by topic, we treat each topic separately and stratify each topic’s data by 10-fold cross-validation to split our training and testing sets. We utilize linear regression with three different regularization methods: LASSO, ridge, and elastic net. Through extensive experiments, we test our model on two datasets (Internet Argument Corpus v2 [1] and Annotated Coarse Discourse [215]), we tested our model on Logistic Regression, Support Vector Machine, and Random Forest and achieved good performance, F-1 scores of over 0.745. Interestingly, our model yielded the highest F-1 score, 0.814, on a stance class that was not taken into consideration in prior work. We report that none of the metrics utilized to measure divergence of opinion yield values exceeding 50% and the correlations between the same topics over 10-fold cross-validation are statistically significant for the majority of them ($p < 0.005$).

For behavioral deterioration, we propose a formal definition of the problem of behavioral deterioration and propose a conceptually simple and highly interpretable method. Our method constructs behavioral sequences from consecutive combinations of misbehavior classes and explores n-gram features to gain a better understanding of behavior exhibited by forum members and predict behavioral deterioration over time. Furthermore, we investigate temporal deterioration patterns from behavioral sequences to predict deterioration at the community level. Through extensive experiments on two datasets (HatebaseTwitter [41] and TRAC [104]), we test our method on Logistic Regression, Support Vector Machine, LSTM [80], Bidirectional LSTM (BiLSTM) [62], and BiLSTM with attention mechanism [27, 193]. Our findings suggest that our method has the potential of leveraging behavioral sequences for predicting deterioration and show the ability of our method in predicting behavioral deterioration with a high degree of accuracy, i.e., F-1 scores of over 0.8. Moreover, we examine the trajectory of behavioral deterioration in order to discover the emotional states that individuals progressively exhibit and assess if these emotional states lead to the degradation of behaviors over time. Our findings suggest that *anger* could be a potential emotional state that can substantially contribute to behavioral deterioration.

Part I

Understanding Emotional States using Psycholinguistic Features

Summary

In this part, we deal with understanding emotional states and moral foundations in the language use of text data. The rationale behind this is to discover whether individuals perpetrating misbehavior manifest social morality and emotional instability. We investigate emotional states and social morality to understand moral differences in a broad spectrum of interactions on social media. Morality guides human social interactions and can potentially conduct to a divergence of opinion, polarity, and hostility when there is moral shock within a community. The key insight is to discover whether differences in moral dimensions have a certain influence on emotional states. To this end, we build a machine learning model using a moral foundations dictionary and propose another model based on natural language inference to automatically extract morality features. We compute the Pearson correlation coefficients between morality and psycholinguistic features extracted from text data to discover the influence of morality on emotions. Furthermore, we examine the temporal evolution of emotional states and identify relevant patterns that lead to emotional instability and breakdown in highly motivated high-conflict interactions such as law enforcement interviews.

In the task of predicting behavioral deterioration, the proposed approaches are crucial because they can help reveal the understanding and involvement of emotional states in the language use of individuals for whom behaviours tend to deteriorate.

Publications

This part has been published as conference papers. Specifically, Chapter 2 has been submitted as “Investigating Moral Foundations from Web Trending Topics” by Jean Marie Tshimula, Belkacem Chikhaoui, and Shengrui Wang for publication at the 25th International Conference on Network-Based Information Systems (NBIS-2022). Chapter 3 is reprinted with permission from “Emotion Detection in Law Enforcement Interviews” by Jean Marie Tshimula, Sharmistha Gray, Belkacem Chikhaoui, and Shengrui Wang. In: IEEE COMPSAC 2022. HCSC: Human Computing & Social Computing.

For the paper (in Chapter 2), Jean Marie Tshimula contributed as the first author. Jean Marie Tshimula designed the proposed model, collected and processed data,

conducted the experiments and wrote the paper. The drafting and verification of the equations were all accompanied by advisors Dr. Shengrui Wang and Dr. Belkacem Chikhaoui.

For the paper (in Chapter 3), Jean Marie Tshimula contributed as the first authors. Jean Marie Tshimula constructed the proposed methodology, collected and processed data, conducted the experiments and wrote the paper. Dr. Sharmistha Gray proposed the utilization the bands in emotion scores to better visualize emotional trajectories. Dr. Shengrui Wang proposed the utilization of the Dynamic Time Warping algorithm to measure the similarity between the emotional trajectories between individuals in order to identify relevant patterns to an emotional breakdown.

Chapter 2

Investigating Moral Foundations from Web Trending Topics

Jean Marie Tshimula,¹ Belkacem Chikhaoui,^{1,2} Shengrui Wang¹

¹Département d'informatique, Université de Sherbrooke, QC J1K 2R1, Canada

²LICEF Research Center, Université TÉLUQ, QC H2S 3L5, Canada
{kobj2801, shengrui.wang}@usherbrooke.ca, belkacem.chikhaoui@teluq.ca

Keywords: Moral foundations, Trending topics, Moral shock, Emotional traits.

Abstract

Moral foundations theory helps understand differences in morality across cultures. Web trending topics assemble diverse opinions on the matters covered in the community. Detecting moral foundations within trending topics-related opinions can be of crucial importance in preventing moral shock and outrage, and extreme actions. In this paper, we propose a model to predict moral foundations (MF) from social media trending topics. Moreover, we investigate whether differences in MF have a certain influence on emotional traits. Our findings show the ability to predict MF, with F-1 scores of over 0.65 and indicate strong evidence that potential

2.1. Introduction

signals relevant to emotional traits can be captured from each MF dimension. Our results are promising and leave room for future research avenues.

2.1 Introduction

Nowadays, media, governments, and organizations keep a constant eye on social media for uncovering the very latest trends. Web trending topics include various opinions on the matters covered in the community. The exploitation of trending topics depends strongly on the goals and interests of each entity. For instance, healthcare organizations may study social media language associated with trending topics to track mental health, symptom mentions and changes in people’s well-being throughout the pandemic [188]. A government may continuously monitor web trending news for identifying rapidly growing divisive and controversial topics that can produce political tensions and an endlessly chaotic situation, and tarnish the country’s reputation and image.

While some trends may involve strong emotional and passionate debate, we need to be clear about the fact that some individuals may be influenced by a wide variety of social and emotional forces and even indulged in emotional manipulation. Research shows that psycholinguistic features and sentiment analysis can provide substantial insights into the issues that affect emotional stability, intensity, and reactions [98]. In this paper, we are interested in examining moral foundations theory to discern moral differences in a broad spectrum of opinions on social media. It is of great importance to understand moral differences at cultural and individual levels because morality guides human social interactions and can potentially conduct to divergence of opinion, polarity, violence, and hostility when there is moral shock within a community. Understanding moral foundations can yield promising results in terms of perceiving the intended meaning of the text data because the concept of morality provides additional information on the unobservable characteristics of information processing and non-conscious cognitive processes [60, 134]. Researchers believe that five psychological dimensions are functioning as foundations for moralities around the world [68].

CHAPTER 2. INVESTIGATING MORAL FOUNDATIONS FROM WEB TRENDING TOPICS

The five moral foundations are care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and purity/degradation. Each dimension possesses virtues and vices:

- Care/harm is associated with the protection of self and others from harm.
- Fairness/cheating evolves through self-interest and reciprocal altruism and is concerned with preserving justice, equity, and trust.
- Loyalty/betrayal is linked to our long history as tribal creatures capable of forming evolving coalitions and is based on the expressions of self-sacrifice for the virtue-vice spectrum, such as patriotism-betrayal, faithfulness-unfaithfulness.
- Authority/subversion underlies virtues of leadership and followership, including deference to legitimate authority and respect for traditions.
- Purity/degradation is associated with sanctity in the virtue dimension and degradation and pollution in the vice dimension.

To the best of our knowledge, our paper is the first to address the problem of moral foundations and emotional traits in web trending topics. In order to address this problem, we utilized Moral Foundations Dictionary (MFD)² and MoralStrength [10]³ in which the five aforementioned psychological dimensions are considered. We combined MFD and MoralStrength and removed duplicate words for each dimension. We trained predictive models on top of zero-shot text classification (ZSC) [210] using the MFD-MoralStrength words as candidate labels.

Specifically, with the scores of candidate labels obtained, we conducted 10-fold cross-validation to separate our training and testing sets and trained two models to predict each dimension: logistic regression and support vector machine. Moreover, we utilized the MFD-MoralStrength words together with the Linguistic Inquiry and Word Count (LIWC) to investigate whether differences in moral foundations have a certain influence on emotional traits.

In line with these contributions, the remainder of this paper is organized as follows. Section 2.2 discusses some related work. Section 2.3 describes the proposed models and data extraction and processing. In all cases, the results we obtain are thoroughly analyzed. Results are extensively discussed in Section 2.5. Finally, Section 2.5 puts

2. <https://moralfoundations.org>

3. <https://github.com/oaraque/moral-foundations>

2.2. Related work

Table 2.1 – Words utilized for prediction. We arbitrarily picked 15 words per dimension from the MFD-MoralStrength dictionary. These words are utilized as candidate labels to feed ZSC for training models to predict each moral foundation dimension.

Care	Harm	Fairness	Cheating	Loyalty	Betrayal	Authority	Subversion	Purity	Degradation
safety	abuse	justice	bias	together	traitor	comply	rebel	innocent	dirty
peace	brutal	right	bigot	loyal	spy	lawful	obstruct	holy	lax
compassion	attack	equity	discrimination	nation	terrorism	respect	protest	abstention	trashy
empathy	hurt	tolerance	exclusion	patriotism	disloyal	order	alienate	limpid	tarnish
care	kill	honest	unfair	commune	enemy	permission	mutinous	maiden	gross
protection	ravage	reasonable	preference	unit	imposter	control	disrespect	virtuous	disgust
amity	ruin	constant	dishonest	devotion	sequester	submission	dissident	decency	pervert
guard	war	unbiased	prejudice	family	miscreant	obedience	riot	pristine	blemish
defense	quarrel	homologous	segregation	group	renegade	leadership	heretic	modesty	profanity
preserve	violence	impartial	favoritism	guild	deceive	duty	defect	immaculate	whore
security	destruction	reciprocal	unscrupulous	fellow	apostate	class	defiance	upright	contagion
shield	crush	balance	disproportion	ally	deserted	position	nonconformist	piety	depraved
benefit	stomp	evenness	inequitable	cohort	individual	hierarchy	oppose	integrity	repulsive
dignity	impair	equivalent	dissociate	member	treacherous	status	denounce	austerity	wanton
refuge	fight	fair	unequal	joint	jilted	authority	remonstrate	pious	sick

forward some concluding remarks and presents future directions.

2.2 Related work

Recent work in trending topics has increasingly focused on structural and semantic features [6, 32], and demographic features [33] for analyzing the contents of the data and its evolution over time. Cheong [32] investigated the topics frequently discussed within an online community and the users’ common interests to understand the meaning and contexts of trending topics. Cheong and Lee [33] proposed demographic information of users to map social media trending topics to understand real-world properties. Althoff et al. [6] processed various trending topics occurred in different periods to analyze the correlation between topic categories and media channels, and to forecast the life cycle of trending topics at the very moment they emerge. Beyond the aforementioned features, our work introduces two features (i.e., morality and emotion) to contribute to the growing body of research seeking to address the problem of trending topics. These features constitute the contribution and novelty of the present paper. The rationale for analyzing trending topics using these features is to understand the moral shock, outrage and conviction, and to prevent moral mandate violations on deviant behavior during trending topics.

To investigate moral foundations in a text corpus, we utilize the Moral Foundations Dictionary (MFD) and MoralStrength. The MFD was developed to operational-

ize moral values and sentiments in the text [4, 55]. Specifically, the MFD provides information on the proportions of virtue and vice words for each moral dimension. MoralStrength was introduced by [10] to improve MFD. One of the advantages of MoralStrength is that it comprises a large number of lemmas and a metric of moral valence for each lemma and provides a moral valence, that is, a quantitative assessment to characterize the lemmas’ relationship with each moral dimension. In this paper, we combined the MFD and MoralStrength to thoroughly analyze moral foundations within trending topics-related text corpora.

2.3 Methodology and data processing

Moral foundations prediction. Zero-shot text classification (ZSC) is a challenging task that aims to make predictions without having seen one single labeled item during training. ZSC associates an appropriate label with a piece of text, irrespective of the text domain and the aspect (e.g., topic, emotion, event, etc.) described by the label. ZSC resembles a human’s ability to generalize and identify new things by transferring knowledge from seen to an unseen domain without explicit supervision [121, 208, 210]. Yin et al. [210] benchmarked zero-shot text classification by standardizing the datasets and evaluations and proposed a textual entailment framework that can work with or without the annotated data of seen labels. Technically, Yin et al. [210] proposed a way to pre-train natural language inference (NLI) models [116] as a ready-made zero-shot sequence classifier. The method works by posing the sequence to be classified as the NLI premise and constructing a hypothesis from each candidate label.

Let $T = \{t_1, t_2, \dots, t_n\}$ be a set of n web posts (such as tweets) from a trending topic and $F = \{f_1, f_2, \dots, f_m\}$ denote moral foundation dimensions, where $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$. Specifically, $c_j = \{w_1, w_2, \dots, w_{15}\}$ designates the fifteen moral foundation dictionary words (see Table 2.1) that fall into each moral dimension f_j . We utilize words related to $\{c_1, c_2, \dots, c_m\}$ as candidate labels input into zero-shot text classification (ZSC) to generate moral scores. Note that the range of moral scores varies from 0 to 1. The values 0 and 1, respectively, indicate the lowest and highest confidence for the relatedness of inputs with labels. The classification of web posts t_i based on each moral dimension can be represented as follows: $S_i^{c_j} = \text{ZSC}(t_i, c_j)$,

2.3. Methodology and data processing

Table 2.2 – Pearson correlation (r) between moral word scores and LIWC features during coronavirus lockdown in Canada.

	Care	Harm	Fairness	Cheating	Loyalty	Betrayal	Authority	Subversion	Purity	Degradation
Positive emotion	0.151	0.014	0.183	0.161	0.246	-0.021	0.053	0.027	0.186	-0.084
Negative emotion	-0.093	0.131	0.054	-0.022	-0.042	-0.030	0.017	0.235	-0.094	0.191
Anger	-0.041	0.128	0.105	-0.019	-0.093	0.191	-0.052	0.176	0.058	0.173
Anxiety	0.038	-0.189	0.033	0.101	0.097	-0.013	-0.006	0.238	0.003	0.219
Sadness	0.195	0.081	0.132	0.004	-0.093	0.155	0.032	0.107	-0.031	0.112

Table 2.3 – Pearson correlation (r) between moral word scores and LIWC features during a specific period of WEXIT.

	Care	Harm	Fairness	Cheating	Loyalty	Betrayal	Authority	Subversion	Purity	Degradation
Positive emotion	0.252	-0.044	0.209	-0.093	0.102	-0.097	0.160	0.122	0.103	0.017
Negative emotion	0.057	0.203	0.073	0.154	-0.046	0.024	0.188	0.001	0.020	0.113
Anger	-0.026	0.145	0.003	-0.059	0.004	0.180	0.028	0.163	0.064	0.154
Anxiety	-0.025	0.152	-0.069	0.271	0.171	0.163	0.166	0.132	0.017	0.150
Sadness	0.283	-0.089	0.152	-0.032	0.168	0.021	-0.051	-0.059	0.258	0.008

where $S_i^{c_j}$ includes the following result $\{s_1^{c_j}, s_2^{c_j}, \dots, s_m^{c_j}\}$, $s_k^{c_j}$ represents the moral score of each word constituting c_j and m denotes the vocabulary size of c_j .

To predict moral foundations from a text corpus, we utilized zero-shot text classification (ZSC), particularly the BART-large-mnli model [116]⁴, to classify moral values in web trending topics. We therefore applied fifteen words for each of ten moral values as candidate labels for classifying trending topics-related text corpora into any of the candidate labels and computing the score for each of the candidate labels (see Table 2.1). We trained logistic regression (LR) and support vector machine (SVM) models by using the scores of candidate labels. For SVM, we set the regularization parameter λ to 0.0001 and the value γ of the radial basis function kernel to 0.5. In order to evaluate the performance of the constructed predictive models, we performed 10-fold cross-validation to split our training and testing sets and computed the F-1 score metric to measure the model accuracy.

Emotional traits. To investigate whether differences in moral foundations have a certain influence on emotional traits, we computed the correlation between MFD-MoralStrength word scores and LIWC features. The LIWC is a widely used, psychometrically validated system for psychology-related language analysis and word

4. <https://huggingface.co/facebook/bart-large-mnli>

CHAPTER 2. INVESTIGATING MORAL FOUNDATIONS FROM WEB TRENDING TOPICS

classification [144]. Specifically, the LIWC comprises word categories that have pre-labeled meanings created by psychologists. The LIWC categories have also been independently evaluated for their correlation with psychological concepts. For each input sequence, we computed the number of observed words, using LIWC and focusing on five LIWC subcategories of psychological processes: *positive emotion*, *negative emotion*, *anger*, *anxiety* and *sadness*. Specifically, we performed the Pearson correlation (r) between MFD-MoralStrength word scores and LIWC features extracted from the text data. To this end, we removed stop words in each sentence and count the number of words. We divided the total number of words that falls either into the virtue or the vice by the total number of words in a sentence. Subsequently, Pearson correlation (r) between the psychological dimensions is applied. The rational behind this is to examine the difference in moral values and moral rhetoric between the five psychological dimensions.

Data extraction and processing. We conducted extensive experiments to investigate two trending topics: coronavirus lockdown measures and the Western Exit (WEXIT) separatist movement in Canada. We collected 857,294 coronavirus lockdown-related tweets posted between 12 March 2020 and 25 May 2020. Specifically, we extracted tweets bearing the words or hashtags: covid, coronavirus, #StayAtHome, or #StayHome. For the WEXIT, we collected nearly 78,000 tweets, posted between 19 Oct. 2019 and 3 March 2020, using the hashtags #wexit and #wexitnow. To preprocess the data, we limited our set to Canada geolocated tweets and removed tweets written in a language other than English or French.

2.4 Discussion

Tables 2.2 and 2.3 show the results of the Pearson correlation (r) between moral word scores and psycholinguistic features extracted from the lockdown and WEXIT dataset, respectively. For coronavirus lockdown, we observe that all correlations between *positive emotion* and virtue moral foundations (*Care*, *Fairness*, *Loyalty*, *Authority*, *Purity*) as well as all correlations between *negative emotion* and vice moral foundations (*Harm*, *Cheating*, *Betrayal*, *Degradation*) are statistically significant at

2.4. Discussion

Table 2.4 – Classification results for moral foundations using the combination of ZSC and predictive models.

(a) ZSC with support vector machine (ZSC+SVM)

	Care	Harm	Fairness	Cheating	Loyalty	Betrayal	Authority	Subversion	Purity	Degradation
Lockdown	0.671	0.673	0.704	0.647	0.688	0.708	0.706	0.710	0.640	0.692
Wexit	0.697	0.630	0.708	0.669	0.697	0.658	0.671	0.693	0.688	0.645

(b) ZSC with logistic regression (ZSC+LR)

	Care	Harm	Fairness	Cheating	Loyalty	Betrayal	Authority	Subversion	Purity	Degradation
Lockdown	0.683	0.655	0.686	0.659	0.703	0.7	0.702	0.687	0.655	0.679
Wexit	0.668	0.642	0.69	0.661	0.705	0.662	0.701	0.676	0.68	0.674

$p < 0.001$; except for one vice moral foundation, *Subversion* ($p > 0.05$). We report that *Degradation* is associated with a relatively high correlation with *negative emotion*, *anger*, *anxiety*, and *sadness* ($p < 0.001$) during the lockdown period. This reflects the feelings that touch directly on mental health.

Research reveals that many Canadians have seen their stress levels double since the onset of the pandemic and are struggling with fear and uncertainty about their own and their loved ones’ health [25]. Survey research conducted by Mental Health Research Canada found that feelings of depression are rising constantly [128]. Before the pandemic, 7% of Canadians reported high levels of depression. This rate has risen to 16% during the lockdown period and 22% predict high levels of depression if social isolation continues for two more months [128]. Our results are alarming and indicate potential signals relevant to mental health that can aid mental health services in assessing the impact of the pandemic on the population and implementing healthier coping strategies to build resilience.

As for the WEXIT, it refers to a movement that advocates for separation from Canada. We report that the correlations between all vice moral foundations and *negative emotion* and *anger* as well as all virtue moral foundations and *anxiety* and *sadness* are statistically significant at $p < 0.001$. We also note that there are some significant correlations between vice moral foundations and *negative emotion*, and virtue moral foundations and *positive emotion* ($p < 0.05$). Our results show relatively low levels of *sadness* for *Harm*, *Subversion* and *Degradation* ($p > 0.05$). We

CHAPTER 2. INVESTIGATING MORAL FOUNDATIONS FROM WEB TRENDING TOPICS

observe that WEXIT conversations include dominant emotional traits for vice moral foundations. This could be considered as strong evidence to argue that this trending topic may have sparked stormy debates and emotional statements. Identifying WEXIT proponents and opponents normally requires further analysis such as stance detection [187], measures of divergent opinions, and community detection to track persistent members of ever-growing communities and their linguistic idiosyncrasies.

Table 2.4 presents the performance results for virtue and vice moral dimensions classification during coronavirus lockdown measures and the WEXIT movement in Canada. We observe that the F-1 scores are higher and show the ability to predict moral foundations, with F-1 scores of over 0.65 for the ten classes. Our model leverages large-scale social media text data stemming from trending topics. The absence of adequate annotated data on moral foundations may be challenging. To overcome this problem, we used psychologically validated and annotated dictionaries that indicate moral foundation dimensions. We applied these resources to track moral foundations using ZSC, a model that does not require data to be annotated beforehand to predict text data; it learns a classifier on one set of labels and then evaluates on a different set of labels that the classifier has never seen before. Note that the lack of annotated data does not affect the generalizability of the findings and model performance.

2.5 Conclusion

We presented the first experiments towards investigating moral foundations from large-scale social media text data from trending topics. Our results provide strong evidence that we can predict moral foundations with an accuracy exceeding 0.65 and we can jointly investigate emotional traits and moral foundations. Though the results are encouraging, this problem leaves room for future research. In the future, we aim to study natural language inference and behavioral deterioration in moral foundations [187].

Chapter 3

Emotion Detection in Law Enforcement Interviews

Jean Marie Tshimula,¹ Sharmistha Gray,² Belkacem Chikhaoui,^{1,3} Shengrui Wang¹

¹Département d'informatique, Université de Sherbrooke, QC J1K 2R1, Canada

²Nuance Communications, 1 Wayside Rd, Burlington, MA 01803, United States

³LICEF Research Center, Université TÉLUQ, QC H2S 3L5, Canada

{kabj2801, shengrui.wang}@usherbrooke.ca, sharmi.gray@nuance.com,
belkacem.chikhaoui@teluq.ca

Keywords: emotion, detection, police, interrogation, transcripts.

Abstract

Understanding the factors that lead or contribute to emotional instability in highly motivated high-conflict dialogues such as law enforcement interviews can be of crucial importance. In this paper, we extract psycholinguistic features to assess emotional stability scale development and identify patterns that are relevant to emotional breakdown. To this end, we utilize zero-shot text classification to investigate the temporal evolution of emotion during law enforcement interviews. We conduct extensive experiments using publicly available police interrogation transcripts. Our results are promising and suggest avenues for future research.

3.1 Introduction

A law enforcement interview (LEI) is a core duty of policing in which the aim is to elicit evidence and accounts from suspects, witnesses, and victims about matters under investigation. The primary goal of information gathering in an LEI is to further the inquiry by establishing facts. The attitudes adopted by the interviewer and interviewee can either facilitate or complicate information gathering. In investigative interviewing, interviewers may encounter interviewees experiencing a range of emotional states that must be accommodated and managed to obtain information about a given event [157]. In psychology, complex states of feeling leading to a change in actions, behavior and personality [189] can be referred to as emotions. Emotions influence perceptions, thinking, motivation and interpretations of events. To manage emotion during LEIs, Risan et al. [157] proposed key considerations, based on the concept of emotional intelligence, that are relevant for interviewer training. The concept introduced by their study emphasizes empathy and emotion regulation. Research has shown that a critical component in the development of rapport in investigative interviewing is empathy, which can be considered as an attitude or a way of relating to experience [82, 125]. Empathy can create an understanding of the other’s experience and perspective, making it possible to show compassion and act based on a perception of the other person’s feelings [17]. As for emotion regulation, Risan et al. [157] defined it as the process of modulating one or more aspects of emotional experience or response; that is, how the individual affects the feeling that is experienced in terms of its intensity, duration, and expression. Emotion regulation concerns how the individual can increase, maintain, or decrease one or more experiential or behavioral components of emotion.

Various factors may lead to negative feelings and reluctance to continue the interview, and understanding these factors can help investigators improve their performance outcomes and take preemptive measures to avoid emotional reactions. Law enforcement interviews take hours and generate large volumes of data in audio and video formats. Listening to tons of interview audio/video to derive actionable insights can be challenging and complex. The copious quantity of LEI transcripts can make it difficult for humans to effectively track emotions exhibited by interviewers

3.2. Related work

and interviewees. Therefore, it becomes necessary to analyze the LEIs to investigate the temporal evolution of emotional stability and assess emotional states during investigative interviewing. This can allow the identification of relevant patterns that lead to an emotional breakdown.

For the work described in this paper, we utilized real-world LEI transcripts which had the advantage of being pre-transcribed to text format from video and audio. Specifically, we extracted psycholinguistic features from text transcriptions to assess scores for emotional aspects such as *pessimism*, *fear*, *anger* and *optimism*. Moreover, we investigated the temporal evolution of emotional states exhibited by interviewers and interviewees, considering the context of past statements expressed by each of them, to identify relevant patterns that can better explain an emotional breakdown during the interview. To the best of our knowledge, our work is the first to utilize a large quantity of real-world LEI transcripts to address the problem of emotion detection in the context of law enforcement interviews.

3.2 Related work

Many studies have recently investigated various aspects of law enforcement interviews such as the language of police interviewing [51, 74, 78, 81]; psychological vulnerabilities during police interviews [63, 135, 142]; cultural considerations [14, 15]; and emotion detection [90, 112, 138, 141, 153, 157].

Psychological vulnerabilities are important in police interviews since they may place interviewees at a disadvantage in coping with the demand characteristics and stress of the interview and in understanding questions and the implications of their answers and statements to police [63]. Regarding the language of police interviewing, research has shown how complex and negative questions may create obstacles in information-gathering and how uncooperativeness can consume precious police time [51]. To fully understand the impact of the language in police interviewing, the authors in [78] and [74] conducted extensive experiments, combining elements of interactional sociolinguistics and critical discourse analysis to demonstrate the need for increased awareness within the criminal justice system of the many linguistic factors affecting interview evidence. It is more likely that interviewees respond emotionally

during the interview. A study by [78] found that when a person becomes emotionally involved, it is easier to be quick-tempered but difficult to return to passivity. In such cases, Risan et al. [157] recommend that interviewers show an awareness of the interviewee’s emotional activation and to express acceptance and understanding of how the interviewee is feeling. More generally, it is always recommended for police to remain impartial, regulate their emotional states and express themselves in ways that will achieve voluntary statements [153]. It is particularly important to measure the feelings of interviewers and interviewees to understand the factors that contribute to emotional instability in their statements during police interviews. Due to the duration of police interviews and the data formats these interviews yield, it is tedious and challenging for humans to manually track the evolution of emotional states exhibited by interviewees and interviewers throughout the interviews. To be able to measure the feelings and other emotional aspects requires a fully automated system based on artificial intelligence.

Our work aims to automate emotion detection in the context of law enforcement interviews and to closely follow the temporal evolution of emotional states during the LEI interview. Beyond our investigation to understand emotional state shifts, we seek to identify relevant patterns that can help to explain emotional breakdowns during police interviews.

3.3 Methods

Zero-shot text classification. Zero-shot text classification (ZSC) is a challenging task that aims to make predictions without having seen a single labeled item during training. ZSC associates an appropriate label with a piece of text, irrespective of the text domain and the aspect (e.g., topic, emotion, event, etc.) described by the label. ZSC resembles a human’s ability to generalize and identify new things by transferring knowledge from a seen to an unseen domain without explicit supervision [121, 208, 211]. The approach by [211] benchmarked zero-shot text classification by standardizing the datasets and evaluations and proposed a textual entailment framework that can work with or without the annotated data of seen labels. Technically, Yin et al. [211] proposed a way to pre-train natural language inference (NLI) models

3.3. Methods

[116] as a ready-made zero-shot sequence classifier. The method works by posing the sequence to be classified as the NLI premise and constructing a hypothesis from each candidate label. One of the advantages for using ZSC is its ability to classify without having received any training examples.

Psycholinguistic features. We extracted psycholinguistic features using the Linguistic Inquiry and Word Count (LIWC) dictionary. LIWC is a widely used psychometrically validated system for psychology-related analysis of language and word classification [144]. LIWC includes words that have very clear, pre-labeled meanings. The LIWC dictionary includes words in various categories: linguistic dimensions, psychological processes and personal concerns. Each category is found to be correlated with several psychological traits and outcomes [65, 66]. We applied zero-shot text classification to classify emotions in each LEI transcription by utilizing the LIWC dictionary and focusing on four LIWC psychological process subcategories: *pessimism*, *fear*, *anger* and *optimism*. Specifically, we considered the LIWC dictionary words that fall into these subcategories as candidate labels and resorted to a zero-shot text classification technique for classifying LEI transcriptions over time. At each timestep, we computed the average of the scores of all candidate labels yielded by ZSC for each psychological factor (that is, *pessimism*, *fear*, *anger* or *optimism*), giving us the average score for the psychological factors expressed during the LEI interview in each timestep. Note that the range of emotion scores varies from 0 to 1. The values 0 and 1, respectively, indicate the lowest and highest confidence for the relatedness of inputs with labels.

Let $T = \{t_1, t_2, \dots, t_n\}$ be a set of n transcriptions in an interview and $C = \{c_1, c_2, c_3, c_4\}$ denote LIWC psychological factors, namely *pessimism*, *fear*, *anger* and *optimism*, where $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, 4\}$. Specifically, $c_1 = \{w_1^p, w_2^p, \dots, w_P^p\}$, $c_2 = \{w_1^f, w_2^f, \dots, w_F^f\}$, $c_3 = \{w_1^a, w_2^a, \dots, w_A^a\}$ and $c_4 = \{w_1^o, w_2^o, \dots, w_O^o\}$ designate the LIWC dictionary words that fall into *pessimism*, *fear*, *anger* and *optimism*, respectively. Since the text transcripts utilized in this work stems from audio and video, we refer to each utterance as a transcription in the datasets. We utilize psycholinguistic information related to c_1 , c_2 , c_3 and c_4 as candidate labels input into zero-shot text classification (ZSC) to generate emotion scores. The classification of transcription t_i

based on each psychological factor can be represented as follows: $S_i^{c_j} = \text{ZSC}(t_i, c_j)$, where $S_i^{c_j}$ includes the following result $\{s_1^{c_j}, s_2^{c_j}, \dots, s_m^{c_j}\}$, $s_k^{c_j}$ represents the emotion score of each word constituting c_j and m denotes the vocabulary size of c_j . We average all emotion scores $s_k^{c_j}$ constituting c_j as follows: $v_i^{c_j} = \frac{1}{m} \sum_{k=1}^m s_k^{c_j}$, where $v_i^{c_j}$ denotes the emotion score of c_j in t_i . This logic is applied to all four psychological factors. Intuitively, we consider the psychological factor yielding the highest emotion value as the emotional state related to the transcription t_i ; that is, $\text{argmax}(v_i^{c_1}, v_i^{c_2}, v_i^{c_3}, v_i^{c_4})$.

3.4 Experiment

3.4.1 Datasets

To empirically measure emotions, we focus on the publicly available police interrogation transcripts of four popular cases: George Huguely, Lee Rodarte, Russel Williams and Bryan Greenwell [115].

George Huguely Case (GHC). George Huguely was convicted of second-degree murder in the killing of Yeardley Love, who also played lacrosse at the University of Virginia and was two weeks away from graduation when she was slain. The murder of Yeardley Love took place on May 3, 2010. Love was found dead in her apartment after Huguely, her ex-boyfriend, kicked a hole in her bedroom door and beat her during a day of heavy drinking, according to trial testimony [115].

Lee Rodarte Case (LRC). Lee Rodarte was a manager at the Bonefish Grill. Rodarte killed his coworker, Savannah Gold, in his car in the restaurant’s parking lot. Rodarte was sentenced to 40 years in prison on March 11, 2021, in exchange for agreeing to plead guilty to second-degree murder, according to the State Attorney’s Office. Other charges of tampering with evidence and abuse of a dead body were dropped in the agreement in Jacksonville, Florida [115].

Russell Williams Case (RWC). Russell Williams, former Colonel in the Canadian Forces, was interviewed by Ottawa Police Detective Sergeant Smyth on February 7,

3.4. Experiment

2010, and ultimately convicted for the murder of two women and sexual assaults on two others [115]. Williams was formally charged with two counts of first-degree murder, two counts each of sexual assault and forcible confinement as well as 82 counts of breaking and entering and attempted breaking and entering.

Bryan Greenwell Case (BGC). Bryan Greenwell shot and killed Jennifer Cain and critically wounded Derrell Wilson on May 13, 2016. Greenwell’s fiancé Jodie Cecil was there when the crime occurred in a Shelby Park apartment, Kentucky [115]. The living victim, Wilson, was crucial in the investigation. When police showed the couple an interview in which Wilson, in poor condition in the hospital, implicates them in the crime, they both admitted their involvement. Bryan Greenwell was found guilty of murder, attempted murder, and tampering with physical evidence and was sentenced to life imprisonment

3.4.2 Data processing

All of the data we used is public and is posted and made available on *Criminalwords.net*. Specifically, we gathered four different police interrogation transcripts from criminal cases occurred in the United States and Canada (§3.4.1). It is important to note that we did not include any data that has been marked as ‘private’ by *Criminalwords.net* or any direct messages. *Criminalwords.net* displays text transcripts of the audio and video and provides case summaries.

Datasets. The datasets GHC, LRC, RWC and BGC consist of 498, 663, 1704 and 202 transcriptions, respectively. The raw datasets comprise 7,120, 9,449, 18,679 and 5,877 words, respectively. We manually corrected misspellings in transcriptions and then removed from the transcriptions text sequences indicating inaudible and nonverbal communication, such as ‘deep inhales’, ‘unintelligible’, ‘background’, ‘pause’, ‘clear throat’ and so on. After removing irrelevant information, we computed the average number of words per transcription. Overall values were 14, 14, 11 and 29 from GHC, LRC, RWC and BGC, respectively. Specifically, we obtained an average of 18 words per transcription for the interviewee (suspect) and 11 for the interviewers (detectives) in

CHAPTER 3. EMOTION DETECTION IN LAW ENFORCEMENT INTERVIEWS

GHC; 13 for the interviewee and 15 for the interviewers in LRC; 9 for the interviewee and 13 for the interviewer in RWC; and 26 for the interviewee and 32 for the interviewers in BGC. These results indicate that the transcriptions consist of short texts. It can be seen that the suspect in RWC appeared to be reserved and laconic, while the suspect in GHC talked more than the detectives and seemed to be relatively verbose and more reactive to questions during the interview.

N-gram analysis. To identify important insights in transcriptions and map out major linguistic features contributing to the progress of the interview, we plotted the top thirty tri-grams in the content of detectives and suspects for each dataset (Figure 3.1 and 3.2). To this end, we removed stopwords in transcriptions but kept wh-question words, possessive adjectives and pronouns. Pronouns can reveal information on people’s emotional state, personality and thinking [144]. For instance, research has shown that individuals susceptible to mental conditions such as depression more frequently use first-person pronouns, suggesting higher self-attention focus [37]. We maintained wh-question words because law enforcement officers and detectives ask questions about the offenses that the suspects are alleged to have been involved in, trying to get the suspect’s version of the story. Discarding the wh-question words from transcriptions would hamper our ability to pinpoint the most asked questions and those that triggered emotional reactions during the interview.

Figures 3.1 and 3.2 show the top-30 tri-grams drawn from the discourse of the suspect and the detectives in each case. We observe an overwhelming number of tri-grams including interjections and exclamations. Note that we retained interjections in the analysis because some researchers consider them part of language [198]; they communicate attitudinal information, relating to the emotional or mental state of the speaker. Examining the results obtained from BGC, we observe that the majority of tri-grams indicate that the suspect spoke optimistically, while also expressing a fear of being followed by people (e.g., *‘kept noticing people’*, *‘noticing people following’* and *‘people following me’*); and we note that the discourse of the detectives contained fear, pessimism and optimism. In GHC, we observe that the suspect discourse exhibited an apparent degree of pessimism and fear (e.g., *‘said murder charges’*, *‘nose started bleeding’* and *‘tell me dead’*); and the detectives displayed optimism and anger. In

3.5. Results

particular, we note that suspects in LRC and RWC displayed anger in their discourse (e.g., ‘*her hit her*’, ‘*uh hit her*’, ‘*put tape her*’, ‘*jeopardizing my job*’, ‘*leave me alone*’, ‘*hanging savannah stuff*’ and ‘*used lot drugs*’); and observe that detectives expressed themselves in pessimistic ways. Overall, we find that the language of the suspects in the four interviews shows they had presumably convinced themselves they were telling the truth (e.g., ‘*i’m pretty sure*’, ‘*i’m serious want*’, ‘*serious want talk*’, ‘*i’m sure uh*’, ‘*i’m saying like*’, ‘*know i’m saying*’ and ‘*i’m saying know*’).

Our thorough inspection above consists of capturing the most mentioned points during the interview to allow for a summary understanding of emotional intelligence. We believe that constant insistence on particular points can reveal the emotional states of the speaker. The reported tri-gram results are an artifact of our data processing: in other words, pronouns, possessive adjectives, interjections, exclamations and wh-question words are significantly more likely in the language of emotional intelligence. The features yielded by tri-grams provide crucial information and offer insights into the conduct of the interviews, allowing us to discern subtle shifts in how answers were given and questions were posed. In §3.5, we measure emotion to identify patterns relevant to an emotional breakdown.

3.5 Results

3.5.1 Psycholinguistic features

We utilized zero-shot text classification (ZSC), particularly the BART-large-mnli model [116]⁵, to classify emotional states in transcriptions. Since ZSC depends highly on candidate labels to measure emotion scores, we considered as candidate labels those LIWC dictionary words that fall into four psychological process subcategories (*pessimism*, *fear*, *anger* and *optimism*). We applied our methodology as described in §3.3 to average the emotion scores for candidate labels of each psychological factor in each transcription and considered the psychological factor yielding the highest average score as the emotional state related to the transcription. Recall that the range of emotion scores varies from 0 to 1. Since the goal of our work is to understand the

5. <https://huggingface.co/facebook/bart-large-mnli>

CHAPTER 3. EMOTION DETECTION IN LAW ENFORCEMENT INTERVIEWS

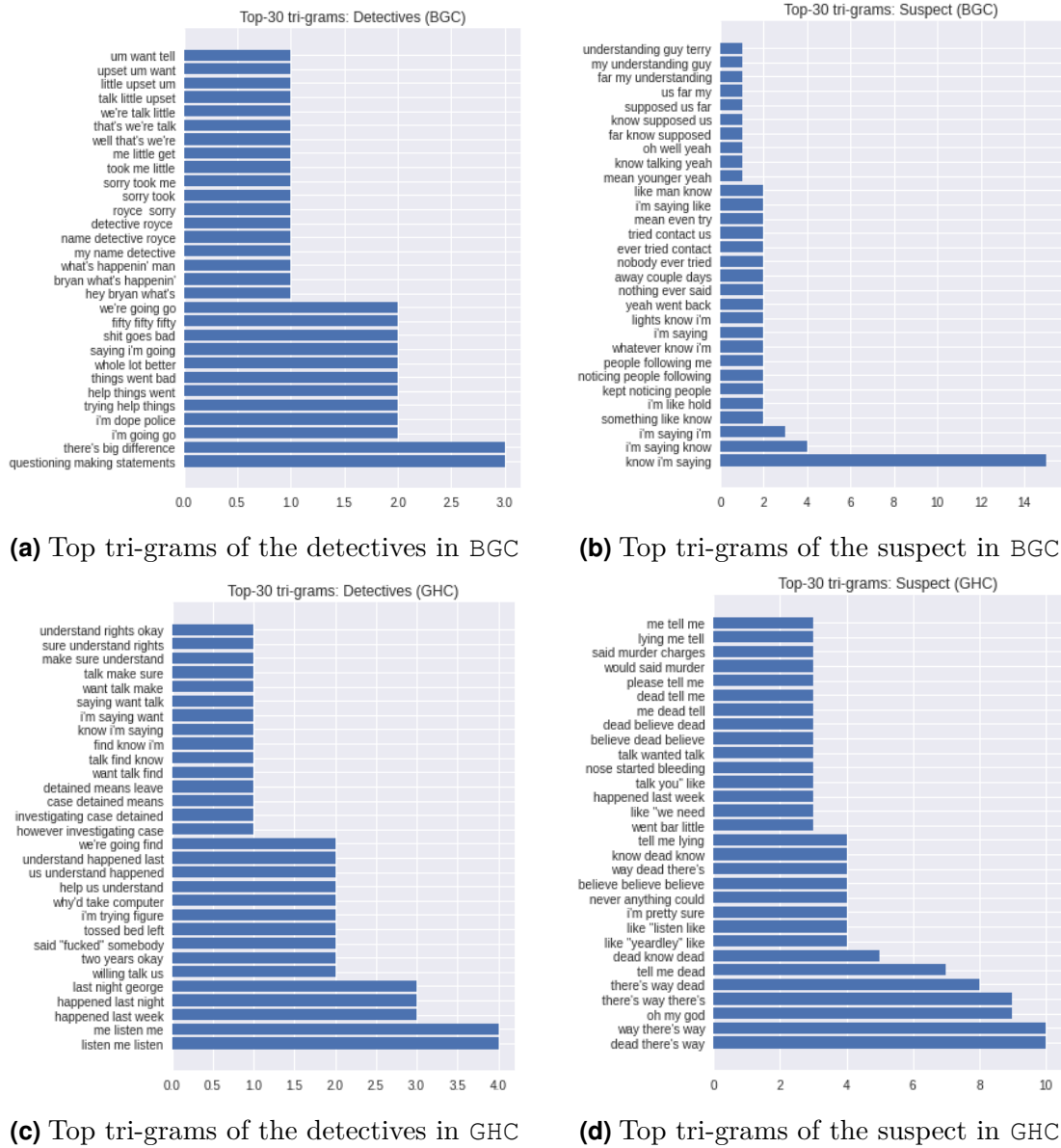


Figure 3.1 – Top 30 tri-grams occurring in BGC and GHC. Note that *x-axis* represents the frequency that a tri-gram appears in the conversation. We observe that (1) the suspect's discourse in GHC displays *fear* and *pessimism*; and (2) the suspect's discourse in BGC shows *fear* and *optimism*.

3.5. Results

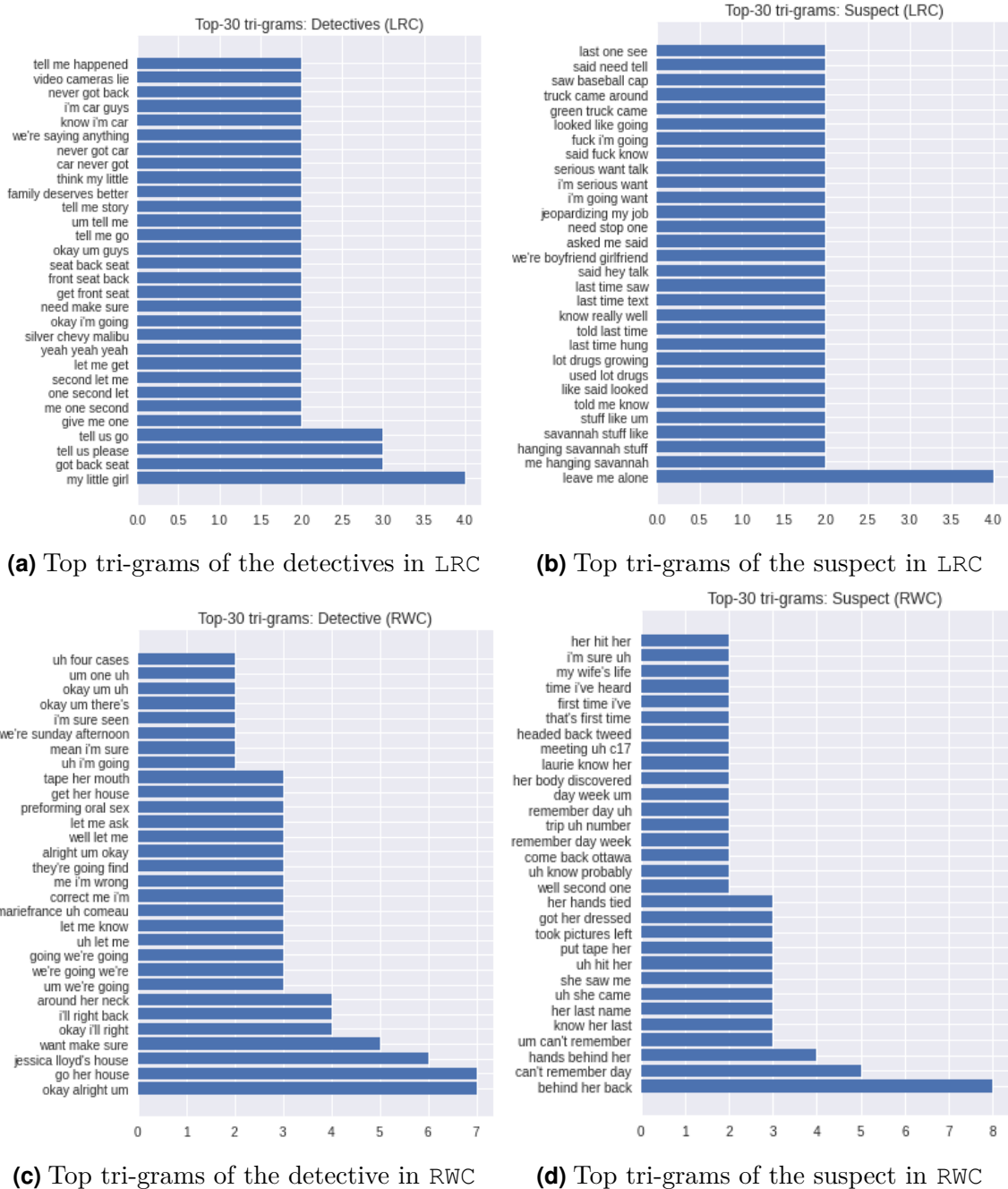


Figure 3.2 – Top 30 tri-grams occurring in LRC and RWC. Note that *x-axis* represents the frequency that a tri-gram appears in the conversation. We observe that (1) the suspects in LRC and RWC show substantial anger-centric content.

CHAPTER 3. EMOTION DETECTION IN LAW ENFORCEMENT INTERVIEWS

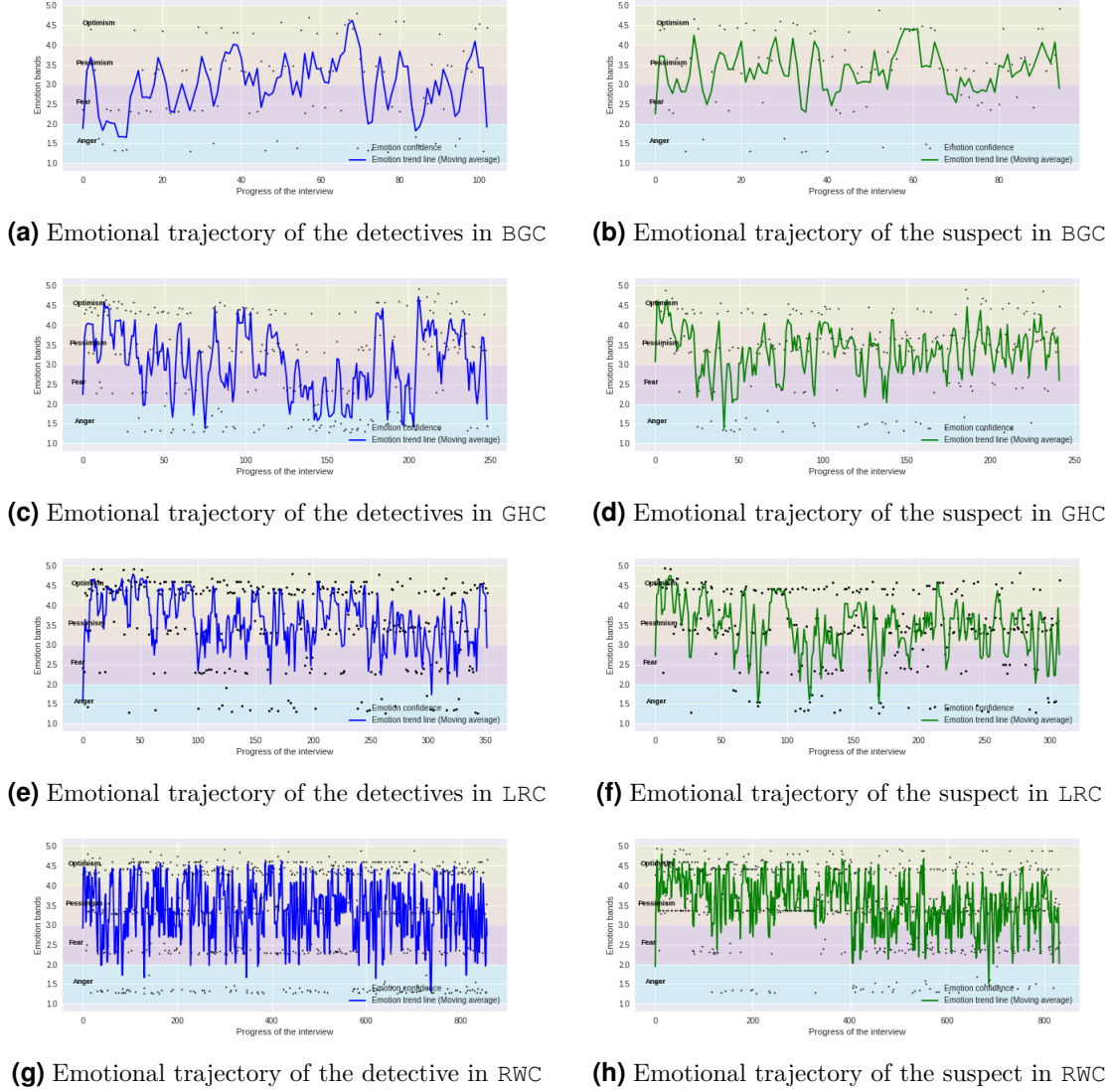


Figure 3.3 – Temporal evolution of emotional states during interviews. Note that *x-axis* represents the index of the transcription as the interview is progressing. Dots indicate emotion confidence scores and colored lines denote emotion trends. We used a window size of 3 to compute a moving average that represents emotion trends.

shifting of emotional states during the interview, we introduced bands in the emotion scores to facilitate the representation of the emotional trajectory of each actor. If we plot the emotional trajectory using emotion scores alone, without banding them,

3.5. Results

the resulting visuals may be hard to read and interpret. One of the advantages of introducing band boundaries is to temporally situate the area (emotional state) into which emotion scores fall. Specifically, we added band values to emotion scores as follows: 4, 3, 2 and 1 for *optimism*, *pessimism*, *fear* and *anger*, respectively. For instance, if a transcription indicates *optimism* and yields an emotion score of 0.78, its banded emotion score is 4.78.

To investigate the temporal evolution of emotion, we considered banded emotion scores and used a window size of 3 to compute a moving average that represents emotion trends.

Figure 3.3 shows the temporal evolution of emotional states during the interview for each criminal case. Based on the relative frequency of emotional states, we observe that the detectives exhibited *pessimism* and *fear* and the suspect displayed *pessimism* and *optimism* the majority of the time in BGC. While the suspect seemed to be more optimistic than the detectives, we find that suspect and detectives expressed their emotional states in approximately the same order of proportion, except for *anger*. In particular, we notice that when the discourse of the detectives contained *anger*, the suspect replied with *pessimism* and *fear*. For GHC, we observe that the suspect's discourse contained a relatively high level of *pessimism*. The detectives showed more *optimism* than the suspect, even if they exhibited a relatively large proportion of *anger* during the interview. The few times the suspect's discourse indicated *anger*, it appeared more likely that he was responding to *anger* with *anger*. For LRC, we observe that the shift of emotional states evolved somewhat at the same pace, even though there is a visible demarcation between the *pessimism* of the suspect and the *optimism* of the detectives. We notice that the suspect's emotions moved back and forth between *pessimism* and *anger* more frequently. For RWC, we observe that detective emotions fluctuated between *optimism* and *anger* and suspect emotions indicated a relatively high proportion of *pessimism*. We remark that the suspect's discourse more frequently contained *anger* after the detective exhibited *anger*.

CHAPTER 3. EMOTION DETECTION IN LAW ENFORCEMENT INTERVIEWS

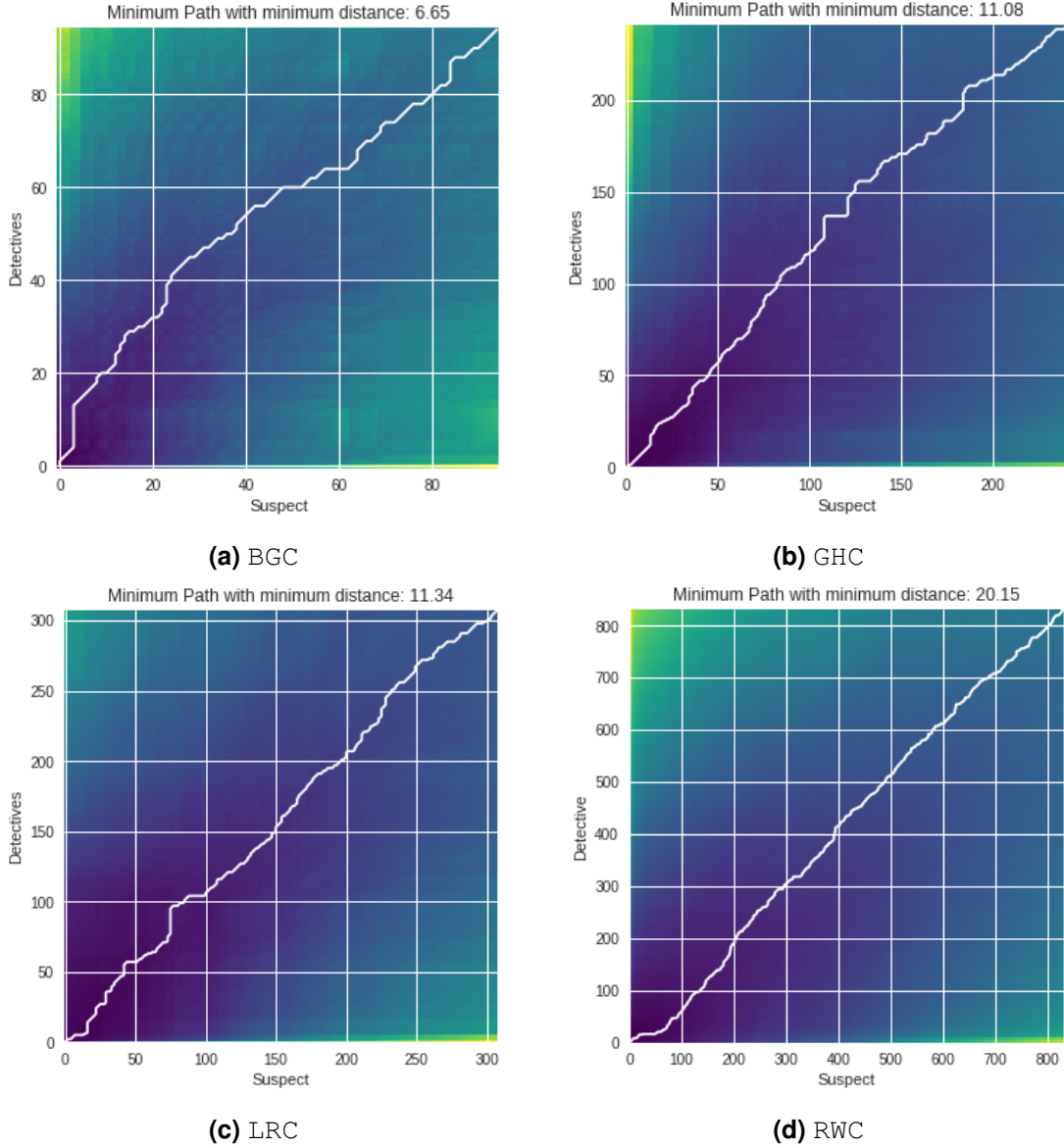


Figure 3.4 – Similarity between the emotional trajectory of the suspect and detectives. The minimum distance of the minimum path is presented at the top of each sub-figure.

3.5.2 Comparison of emotional trajectories

We calculated Pearson's correlation coefficients (r) and utilized the Dynamic Time Warping (DTW) algorithm [88] to measure the similarity between the emotional

3.6. Ethical considerations

trajectories of the suspect and the detectives in each case. DTW is a very robust technique that computes the path between two temporal sequences which minimizes the distance between them. More interestingly, DTW allows non-linear alignments, handles sequences of different lengths, and calculates the euclidean distance of each frame from every other frame to compute the minimum path that will match the two temporal sequences. The closer the distance value is to 0, the more similar the two temporal sequences are. The motivation for utilizing DTW is that our datasets include multiple suspect transcriptions that are related to a particular detective transcription and vice versa.

We obtained low correlations between emotional trajectories in all of the criminal cases. The Pearson correlation coefficients for LRC ($r=.303$, $p<.001$) and GHC ($r=.1784$, $p<.05$) are statistically significant, indicating that suspect and detectives exhibit certain patterns that look relatively similar. In addition, we report the correlations ($r=.1361$, $p>.05$) for BGC and ($r=.038$, $p>.05$) for RWC, which are not statistically significant.

Figure 3.4 shows the similarity between the emotional trajectories of the suspect and the detectives. It can be seen that the minimum distance yielded a value of 6.65, 11.08, 11.34 and 20.15 for BGC, GHC, LRC and RWC, respectively. We observe that BGC achieved the lowest minimum distance. This suggests that the emotional trajectories of the suspect and the detectives in BGC includes a relatively high proportion of similar patterns. Note that some pairwise comparisons yield statistically significant correlations of small magnitude in a positive direction. Taken with the previous results, this shows that relationships between suspect and detective emotions can exhibit individual-level variability in both direction and magnitude and that inferences about these relationships must explicitly and quantitatively account for this variability. Furthermore, this highlights the fact that significant signals remain to be uncovered and understood in the language of police interrogation.

3.6 Ethical considerations

The sensitive nature of criminal investigation research requires us to consider possible benefits of this study. The potential immediate benefit of this study is the

CHAPTER 3. EMOTION DETECTION IN LAW ENFORCEMENT INTERVIEWS

extraction of psycholinguistic features to assess emotional stability scale development. A potential secondary benefit is the identification of relevant patterns to explain an emotional breakdown during police interviews. The results presented here demonstrate the feasibility of using automated machine learning to investigate the temporal evolution of emotional states during law enforcement interviews.

Aside from our interest in investigating emotional states in police interrogation transcripts, we do not exhibit sociodemographic characteristics of suspects, nor contest or comment on the police interrogation approach utilized by detectives or the conviction judgments in each case. Our work takes a step toward implementing a system that automatically examines police interrogation transcripts and monitors emotional thresholds in order to ensure that detectives and suspects are emotionally stable and predict whether they tend to react emotionally.

3.7 Discussion and conclusion

This work focuses on measuring emotions from police interrogation transcripts, investigating the temporal evolution of emotional states, and identifying relevant patterns that may signal emotional breakdown. Our data processing demonstrates that signals relevant to emotions can be extracted through n-gram analysis, and finds that tri-grams containing interjections more frequently indicate *optimism*. We measure the scores of emotional states, utilizing four LIWC psycholinguistic features (*pessimism*, *fear*, *anger*, and *optimism*). Our findings indicate emotional trajectories illustrating shifts in emotional states and show similarities and correlations between the emotional trajectories of suspects and detectives. We found that our datasets include a relatively high proportion of patterns.

While the results reported here hold promise for future work, both theoretical and applied, our research is limited by several important factors. We believe that there are still significant signals that must be investigated and considered to build a holistic model for emotion detection in the context of law enforcement interviews. For instance, the audio and video format of police interrogation transcripts include tones of voice, body language and facial expressions. This study did not take into consideration the aforementioned features. In future work, we would like to (1) build

3.7. Discussion and conclusion

an entire pipeline for automating emotion detection process in police interrogation in real-time, (2) examine behavior deterioration from emotional trajectories [187] and (3) compare suspect patterns with each other to construct profiles and examine the relationship between emotional intelligence and criminal behavior [164]. We would also like to identify inherent biases in police/detective mind and train a BERT model [43] for police interrogations using transcriptions involving serious felonies and light crimes.

Our results identify several opportunities and challenges in the development of machine learning tools that rely on various features including body language and facial expressions to examine police interrogation transcripts.

Part II

Discovery of Affinity and Personality from Text Data

Summary

In Chapter 4, we address the challenge of discovering affinity relationships between social media users. The problem of affinity relationships in the context of social media has not been clearly and formally defined in the literature. We propose mathematical definitions of affinity influence and extract several interpretable features from these definitions. The challenge of discovering affinity goes beyond carrying out an analysis based on structural features such as *likes*, *shares*, *followers* and *followings*. Rather than relying solely on structural features of social media, we combine structural features, temporal information and the content of interactions. We develop an advanced method based on Markov models, machine learning and natural language processing to quantify affinity scores using the combined features. We utilize the quantified affinity scores to investigate the evolution of affinity over time and predicting affinity relationships arising from the influence of certain users. We show that our approach achieves good performance compared to state-of-the-art techniques, and results in robust discovery and considers minute details.

In the task of predicting behavioral deterioration, the proposed approach can reveal individuals who seem to foment misbehavior in social media platforms and assess the likelihood that their relationships can evolve and the risks they may represent.

In Chapter 5, we investigate the influence of personality on affinity. The rationale behind this is to discover affinity relationships between different personality types. The combination of affinity and personality allows us to understand how individuals with similar personality traits get to develop their affinity and discern what attracts an individual to another. In contrast to psychological research, we utilize the language use of text data to evaluate personality and emotional states; we propose approaches that utilize psycholinguistic features, to measure emotional states and understand their linguistic idiosyncrasies. More specifically, we derive affinity relationships between individuals using the approach developed in Chapter 4 and examine personality based on the language use to discover the emotional stability of affinity relationships, and measure semantic similarity at the personality type level to understand the logic behind the development of affinity. The key motivating insight is that our results identify certain influential personality types that weigh more heavily on affinity relationships and show that personality can be predicted from the spontaneous language

with an F-1 score superior to 0.76. In addition, we find that semantic similarity and emotional (in)stability constitute an important lead for understanding the implications of personality in the development of affinity. Our results identify a number of statistically significant correlations in terms of emotional stability in personality-based affinity relationships. The theoretical and practical implications of our outcomes can be valuable for supporting decision-making processes in various domains, including clinical psychology, forensic psychology, digital forensics, human factors and social science. In reality, investigations into the influence of personality can be driven by the concrete needs of applications; for example, the role that personality plays in the effective functioning of behavioral deterioration.

Publications

Chapter 4 has been published as a conference and journal paper, respectively: (1) “HAR-search: A method to discover hidden affinity relationships in online communities” by Jean Marie Tshimula, Belkacem Chikhaoui, and Shengrui Wang. In: Proceedings of 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 176–183 (2019); and (2) “A new approach for affinity relationship discovery in online forums” by Jean Marie Tshimula, Belkacem Chikhaoui, and Shengrui Wang. In: Social Network Analysis and Mining, volume, 10, 40. Specifically, the journal version is provided in Chapter 4.

In these papers, Jean Marie Tshimula contributed as the first author. Jean Marie Tshimula designed the proposed model, collected and processed data, conducted the experiments and wrote the papers. The drafting and verification of the equations were all accompanied by advisors Dr. Shengrui Wang and Dr. Belkacem Chikhaoui.

Chapter 5 has been submitted for publication. Specifically, This research has been submitted as “Discovering Affinity Relationships between Personality Types” by Jean Marie Tshimula, Belkacem Chikhaoui, and Shengrui Wang for publication at the 25th International Conference on Network-Based Information Systems (NBIS-2022).

In this paper, Jean Marie Tshimula contributed as the first author. Jean Marie Tshimula built the proposed methodology, collected and processed data, conducted the experiments and wrote the paper. The proofreading and verification of the equations was accompanied by advisors Dr. Shengrui Wang and Dr. Belkacem Chikhaoui.

Chapter 4

A New Approach for Affinity Relationship Discovery in Online Forums

Jean Marie Tshimula,¹ Belkacem Chikhaoui,^{1,2} Shengrui Wang¹

¹Département d'informatique, Université de Sherbrooke, QC J1K 2R1, Canada

²LICEF Research Center, Université TÉLUQ, QC H2S 3L5, Canada
{kabj2801, shengrui.wang}@usherbrooke.ca, belkacem.chikhaoui@teluq.ca

Keywords: Affinity relationship discovery, Online communities, Online discussions, Natural language semantics, Sentiment analysis, Embeddings.

Abstract

The concept of affinity relationship discovery is relatively new in the context of online discussion communities and there has been little work addressing it to date. This problem entails finding affinity relationships in a community by combining structural features and the content of interactions. Affinity discovery seeks not only to identify these affinity relationships, but also to quantify them so that the degree of affinity between individuals can be perceived in the form of a score. This paper proposes an algorithm based on

CHAPTER 4. A NEW APPROACH FOR AFFINITY RELATIONSHIP DISCOVERY IN ONLINE FORUMS

Markov chain models, named HAR-search, for discovering hidden affinity relationships and deriving affinity scores between individuals in an online community. We demonstrate that our method is capable of tracking the evolution of affinity over time and predicting affinity relationships arising from the influence of certain community members. Comparison with state-of-the-art methods shows that our method results in robust discovery and considers minute details.

4.1 Introduction

The number of online social networks (OSNs) has dramatically increased and each of them has some features that make its business model original and unique. OSNs serve a wider range of purposes than anticipated. Apart from instant messaging and audio and video calls, they now allow the sharing of files and locations as well as the use of many other interesting features. For some, they are a place to have fun, for others, a goldmine to explore. For instance, companies can use OSNs to post ads and conduct campaigns and surveys concerning their products and services, because OSNs offer valuable data about customers. Governments can exploit them by performing advanced sentiment analysis to discern essential viewpoints embedded in unstructured data and succinctly grasp what people are saying about the country. In addition, forensic investigators can use OSNs as a path for conducting investigations, monitoring suspicious hashtags or information, unmasking wanted individuals or fugitives, and decrypting coded (riddle) messages that might herald terrorist activity.

Message boards and messaging services offer a very convenient chat room where people talk about personal experiences in privacy [69]. Basically, such one-to-one communication is not open and does not give access to third parties. On the other hand, group discussions on OSNs link and bring together many people, both unknown and known. Here, people express their opinions and befriend others, especially when they have similar viewpoints. An affinity relationship can be detected in the communication when individuals have various things in common, such as interests, viewpoint, etc., or by the way that individuals exchange. As long as they keep exchanging posi-

4.1. Introduction

tively, this can considerably reinforce their affinity relationship. People progressively develop affinity, and get closer as they mention each other frequently in messages and send positive messages to one another. Depending on the objectives set by a given group, the group's purpose may revolve around topics like science, politics, and sports. In many group discussions, there are people who are reluctant to engage, not because they lack compelling points of view but because they do not have a strong affinity with other community members. Members whose affinities are solid could piggyback themselves over stormy discussions and provide supporting arguments to protect themselves from humiliation [170, 184].

Affinity discovery may play a major role in domains such as (i) recommender systems, by suggesting items rated and preferred by a user to her friends based on their affinity score; (ii) advertising campaigns, by personalizing users' ad content based on the shopping experience of friends with whom they have affinity; (iii) police investigation, by asking individuals who have an affinity with suspects to provide any information that may lead to potential solutions, etc. The affinity score between two community members can be derived from the number of times that one responds to the other's messages, the number of times that one tags (or mentions) the other in messages and the number of messages with positive connotations that one addresses to the other in online discussion forums.

For example, WhatsApp offers one-to-one, one-to-many and group communications [163]. In some WhatsApp-like and message board platforms, users usually mention each other using the @ symbol followed by the username. On Twitter, the affinity score can be derived from the number of times a given user has identified another in their tweets. Mentions also affect the relationship score between the sender and the mentioned person. Mentioning (or tagging or identifying) a person in a message can be considered as a sign of affinity or it can affect the affinity if the message content seems useful and/or informative to the intended person, as well as if the sentiment of that message is positive.

Rezgui et al. [156] introduce a method of affinity discovery using Twitter, although its data is much less structured and dialogical. This method considers mentions only and performs sentiment analysis to derive the affinity. However, this method presents several limitations: (1) it does not capture the context of the messages based on their

CHAPTER 4. A NEW APPROACH FOR AFFINITY RELATIONSHIP DISCOVERY IN ONLINE FORUMS

time series order and the flow of the discussion; and (2) it does not explore affinities over time to learn about their evolution. These limitations impede its ability to capture minute details. In contrast, our method takes into account the order of messages and their context in the discussion. Additionally, we do not remove emojis but convert them into their textual equivalents.

In this paper, we propose a search-based method for discovering affinity relationships between individuals in an online discussion community. This method entails deeply analyzing the online discussion data (ODD) by considering the flow of the discussion to understand the context of messages, in order to discover the degree of affinity that one may have for others, and also to monitor users capable of influencing other community members to enter the discussions and quietly share their opinions (stances or ideas).

Specifically the main contributions of this paper can be summarized as follows:

- We propose an algorithm to discover hidden affinity relationships in online communities.
- We track the evolution of the affinity between actors over time to predict affinity relationships arising through the influence of certain community members.
- We conduct extensive experiments using real datasets to validate the hidden affinity relationships discovered and their evolution over time.

The rest of this paper is organized as follows: A brief outline of some related work is given in Section 4.2. Section 4.3 describes the proposed method. The data is presented in Section 4.4. We discuss the results obtained in Section 4.5. Finally, we conclude and present future directions in Section 4.6.

4.2 Related work

The detection of affinity relationships is relatively new, and only a few papers have addressed this problem to date, in domains such as recommender systems [83, 100], immigration [77], coalitional games [19, 169], and social-economic systems [123]. Sawhney et al. [160] introduced a natural language processing (NLP)-based method to capture details about user interactions and understand structure and semantic information from texts. On similar lines, Rezgui et al. [156] proposed AffinityFinder,

4.2. Related work

a tool with the capacity to extract user tweets containing mentions in order to derive affinity relationships and generate the affinity graph. In particular, this tool uses sentiment analysis to capture the affinity relationships between pairs of users. The limitations of AffinityFinder concern the ability to differentiate retweets from genuine tweets. Since both types contain mentions, the tool treats retweets as if they were normal tweets. Moreover, AffinityFinder is not able to trace the course of a tweet from its seed to retweets and replies, and thus disregards the context of tweets. The authors have not indicated how they measured affinity from tweets, replies and retweets with comments. It should be mentioned that, in the literature, the AffinityFinder method is the first attempt to tackle the problem of hidden affinity detection in the context of online discussion communities, although the problem was not well formulated to some extent. From the discussion forum perspective, affinity detection should basically be considered as the combination of structural features, temporal information, and content of interactions. Note that affinity detection is different from the problem of discovery of implicit or hidden relations addressed by [2, 13, 93, 114, 149, 173, 180, 204, 206, 218]. Most of this work relies exclusively on the structural properties of social networks. As mentioned in Section 4.3, affinity relationships are measured from certain characteristics derived in the discussion content.

To measure the social affinity of individuals from the same social circle, Panigrahy et al. [140] suggested a measure that combines the shortest-path distance between a pair of users and the number of edge-disjoint paths between them. Hong et al. [83] calculated movie similarity from the movie preferences of members in a group. Given that the measure relies on metrics such as the number of interactions between individuals, it may fail to capture the set of components that may best explain the affinity. For example, two users may appear to be closer to one another although they always have some divergence of opinions and do not show mutual appreciation. Say that two users often get into stormy discussions in the forum [152]: the data history may give the impression of closeness, while the content may indicate the opposite. To overcome this problem, we therefore propose to incorporate NLP-based methods for measuring affinity by making use of the sarcasm, temper, intent and moods that users express in text messages while discussing. To learn relationships between sentences, we use BERT [43], a contextual pre-training method that exploits

CHAPTER 4. A NEW APPROACH FOR AFFINITY RELATIONSHIP DISCOVERY IN ONLINE FORUMS

a masked language modeling objective to enable the training of bidirectional models and also adds a next sentence prediction task to improve sentence-level understanding. BERT receives pairs of sentences as input and learns to predict whether sentence B is the actual sentence that comes after sentence A or merely a random sentence. BERT performs next sentence prediction, but does not specify the sentiment of the next sentences obtained.

To analyze semantic and syntactic relations between words in sentences (messages) sent by users, we use word embeddings (`word2vec`) [129]. Fundamentally, affinity can be positive or negative to some extent—for instance, if someone is always taking an opposite viewpoint from someone else in discussion. Negative affinities may be built by sentences expressed respectfully to contradict someone without insults, etc. Such interactions are taken as positive based on the context in which they occur. In this work, we limit ourselves to detecting affinities in an overall way without distinguishing whether they are positive or negative. On the other hand, we assume that only positive sentiments can generate affinity. Our algorithm uses BERT and word embeddings to capture the context of messages by following the flow of the discussion to determine the sentiment of messages, then models positive interactions in the form of sequences to ease the discovery of affinities. To compute the affinity scores, we resort to Markov chain models, because they are more flexible, combine all transitions in one matrix, and calculate more than one-step transition probabilities [168].

4.3 Proposed method

The concept of hidden affinity relationships in social networks has not been clearly and formally defined in the literature. In this paper, we define affinity relationships as being relationships that include a set of characteristics such as mutual understanding, reciprocal and common interests, sympathy, harmonious communication or agreement between individuals. Hidden affinity relationships are such relationships that are more subtle and difficult to observe among many others and are concealed in the midst of the community [185]. The more the interactions of the community increase, the harder it is to detect these relationships.

4.3. Proposed method

In this section, we present the essential steps of our method. To discover clues to affinity, we collect messages sent by each community member, including messages which were addressed to her by other community members. We first use BERT and word embeddings to analyze relationships across words to determine the sentiment by following the context of messages throughout the discussion history: that is, we consider the previous messages to understand the context of the current message. We compress the discussion into sequences with positive interactions, then use Markov chain models to discover hidden affinities and quantify their scores.

4.3.1 Sentiment filtering

Sentiment analysis, also known as opinion mining, is a field of NLP that identifies, extracts and classifies opinions from the text. Sentiment classification algorithms seek to identify whether the text contains a positive or negative opinion (or sentiment) towards the topic. To determine the sentiment of messages, we first label our datasets (Table 4.2) using the unigram dependency-based approach to sentiment analysis [45, 46] and the sentiment lexicons produced by [24, 71, 85]. We record 1 when the sentiment is positive and 0 when the sentiment is negative. Word2vec uses word embeddings to capture the context of words and produce high-dimensional vectors in a space [129].

We have opted to use this tool to build a sentiment classifier, given that word2vec learns vector representation of words. To produce a distributed representation of words, word2vec utilizes one of two models: continuous bag-of-words (CBOW) or skip-gram. CBOW predicts target words from context words and skip-gram predicts context words given the target words. We consider emojis as part of messages, whereas the existing method [156] removes them like stop words.

To learn a better quality of word embeddings, we initialize the embedding with the CBOW model trained on the labeled Google News corpus (about 100 billion words) containing 300-dimensional embeddings for three million words and phrases. This yields the vector representation of each word forming the message. We combine these vectors together to represent the message as a whole. We calculate a weighted average of these vectors, where each weight provides the significance of the word

Algorithm 1 HAR-search

Input : D

Output: backward/forward information for each message

```

1:  $dataset \leftarrow null$ 
2:  $D_t \leftarrow D(d)$ 
3:  $ActorList \leftarrow distinct(D_t(a_i))$ 
4: for  $actor$  in  $ActorList$  do
5:    $rows \leftarrow D_t(actor)$ 
6:   for  $r$  in  $rows$  do
7:      $backward = Previous(r, D_t(s_r))$ 
8:      $forward = Next(r, D_t(s_r))$ 
9:      $dataset.add(actor, backward, forward)$ 
10:  end for
11: end for=0

```

vis-a-vis the message. To classify messages as positive or negative, we use the new representation of messages to train logistic regression and random forest classifiers. Recall that the messages are already labeled. We have adopted the logistic regression classifier since it outperforms the random forest classifier in terms of accuracy (note that this concerns only the accuracy of the sentiment classifier in Section 4.4).

4.3.2 Algorithm of HAR-search

As stated above, AffinityFinder does not determine the sentiment of messages by verifying the context in which they are written. This hampers its ability to consider all details in deriving the affinity score. The sentiment analysis approach used in this method is based on unigram dependency. The unigram dependency approach overlooks the context, the grammar and the order of words in a sentence [45, 46]. HAR-search intends to understand the context of messages with respect to the discussion history, and it identifies positive interactions and models them in sequences in order to discover hidden affinities and determine their scores.

Formally, our algorithm is described as follows: suppose a dataset D that includes a set of variables (d, A, M) , where d represents dates (or period); $A = \{a_1, \dots, a_K\}$ denotes the set of actors (or community members), where K is the number of distinct

4.3. Proposed method

Table 4.1 – An example to illustrate the functioning of HAR-search.

(A) Dataset simulation				
Id	Date	Actor	Message	Label
1	11/11/17, 06:31:19	A	Hey, where can I buy new shoes?	1
2	11/11/17, 06:31:45	D	You are stupid	0
3	11/11/17, 06:34:42	B	Go to MyShoe shop at downtown	1
4	11/11/17, 06:35:11	A	Thank you very much	1
5	11/11/17, 06:37:33	C	Hello @B, it has been a long time	1
6	11/11/17, 06:41:07	B	You are welcome @A. Hi @C, what's up?	1
7	11/11/17, 06:42:53	A	In case, the shop is closed. Could u suggest another one?	1
8	11/11/17, 06:43:16	B	I am doing well despite some fever	1
9	11/11/17, 06:55:23	D	Get well soon	1
10	11/11/17, 06:58:01	C	@A u can try M2shoe near OpenPiz. @D LMAO	1
11	11/11/17, 06:03:39	A	Thanks a lot	1
12	11/11/17, 07:04:14	D	I hate you man, you are an idiot!!!	0

(B) Actors		(C) Tidy dataset			
#	Actor	Id	Actor	Backward	Forward
1	A	1	A		B
2	B	4	A	B	C
3	C	7	A	B	B
4	D	11	A	C	
		3	B	A	A
		6	B	C	A
		8	B	A	D
		5	C	A	B
		10	C	D	A
		9	D	B	C

(D) Positive interaction sequences based on HAR-search	
Id	Sequence
1	A-B
4	B-A-C
7	B-A-B
11	C-A
3	A-B-A
6	C-B-A
8	A-B-D
5	A-C-B
10	D-C-A
9	B-D-C

members in the community C ; and $M = \{s_1, \dots, s_L\}$ denotes sentences (or messages) sent in C , where L is the number of messages in C . We use the feature “backward” to list all actors that have responded or addressed a message to a given actor in the previous discussion lines, and the feature “forward” to list actors that have responded to the actor’s message in the following discussion lines.

Initially, our algorithm is composed of several steps: (1) mention detection in messages [101]; (2) conversion of emojis to text [64, 131, 201]; and (3) analysis of

CHAPTER 4. A NEW APPROACH FOR AFFINITY RELATIONSHIP DISCOVERY IN ONLINE FORUMS

sentiments. In practice, we detach the first task from the remaining ones. We store in a vector all mentions detected in a message. When the sentiment of that message is positive and the vector is not empty, the mentioned users are returned either as users the actor has responded to or as users to whom the actor has addressed the message.

D_t implies a subset derived from D by filtering a particular period d , and then sorting occurrences by timestamp in ascending order, denoted by $D(d)$. `ActorList` selects a distinct list of actors in D_t . The first step consists of identifying all the messages sent by each actor. The second collects intended information and analyzes it with respect to previous and following messages; we apply BERT to identify all possible next sentences that each sentence might have. We use `Previous` to examine occurrences between row r (current message) and row $r-1$ (previous messages). $r-1$ is the single previous message, $r-n$ for $n \geq 1$ is the set of previous messages. Since 92% of the online population use emojis in their communications to voice their stance and opinion [64, 131], we do not remove Unicode characters corresponding to emojis in messages. Having said that, we use sentiment filtering (Section 4.3.1) to determine whether the current message responds to previous messages that have been addressed to the related actor or previous messages with positive sentiment in which the related actor was tagged. Basically, positive sentiment messages do not imply antisocial behavior, including flaming, offensive language, bullying, hate speech, profanity, and insults [31, 75, 86, 154, 155]. They can however imply antisocial behavior such as grooming, manipulation, fraud, narcissism, etc.

`Next` (in Algorithm 1) scrutinizes occurrences between row r and row $r+1$ (following messages). $r+1$ is the single next message, $r+n$ for $n \geq 1$ would be the set of following messages. Here, our goal is to identify messages that respond to the current row message with positive sentiment, including messages containing positive emoji(s) that express happiness (or agreement, satisfaction, etc.) as well as messages containing mention(s) with positive sentiment. The output is a tidy dataset with only three features (`actor`, `backward`, `forward`). These features are further used to model sequences in order to derive the affinity scores.

Table 4.1 shows the abstract of a simulated discussion to illustrate how HAR-search works. This toy example includes an ODD of 12 rows of messages in which

4.3. Proposed method

4 actors are involved in the conversation, as depicted in Table 4.1-A. Table 4.1-B presents the list of actors who are part of the forum. Table 4.1-C summarizes the course of the discussion for each observation by indicating the actors in the previous and next rows who affect the affinity with the actor being processed. This yields a tidy dataset for which we know the values of the backward and forward features. We use the sequence “backward+actor+forward” as described in Table 4.1-D to model the Markov chains in order to calculate the transition matrix, each of whose elements represents the affinity score between a pair of community members.

The order of rows $r-1$, r , and $r+1$ is not sequential in the dataset. For example, the affected rows (r) that contain the positions of actor A are 1, 4, 7 and 11. Consequently, if we want to verify the backward and forward features from row 7, backward $r-1$ points at row 4 and forward $r+1$ implies row 11. As we can see in Table 4.1-A, actor A has addressed her request to everyone in the chat room; B and C have been helpful by providing information appropriate to A ’s request.

4.3.3 Deriving affinity scores

In this paper, we restrict ourselves to the discrete-time case. We introduce first-order Markov chains as the target concept for mining the positive interaction sequences (PIS). The PIS align the names of actors that have participated in positive interactions. A first-order Markov chain is a discrete-time process for which the future behavior of the system depends only on the current state and not on the previous states. Let X be a sequence of random variables $\{X_n\}$ representing the PIS generated by HAR-search (e.g., Table 4.1-D), and let S be a set of states representing actors. $\{X_n\}$ is a Markov chain if it satisfies the Markov property: for any positive integer n and possible states $i, j \in S$,

$$P\{X_{n+1} = j | X_n = i, \dots, X_0 = i_0\} = P\{X_{n+1} = j | X_n = i\} \quad (4.1)$$

$$p_{ij} = P\{X_{n+1} = j | X_n = i\} \quad (4.2)$$

$$P = (p_{ij}) \quad (4.3)$$

The parameters of a first-order Markov chain can be defined by a transition ma-

CHAPTER 4. A NEW APPROACH FOR AFFINITY RELATIONSHIP DISCOVERY IN ONLINE FORUMS

trix P with matrix elements (transition probabilities) p_{ij} , where p_{ij} represents the conditional probability for a transition from state i at time n to state j at time $n+1$. Since p_{ij} is a probability, it must be a value between 0 and 1. Recall that the rows of any state transition matrix must sum to 1. When the sum of any row is equal to 0, the elements constituting this row are considered to be dangling nodes, i.e., nodes having no outgoing links [12]; these were eliminated in practice.

In this section, the novelty of our approach is the way it deals with the PIS to compute the transition probability (p_{ij}). It can be seen that the size of the PIS varies significantly when a message does not have any link with preceding ones or a message has not been answered by other actors. The maximum length of a PIS is 3 (backward, actor, forward). We may also have some PIS of length 2 (see Table 4.1-C and Table 4.1-D): i.e., the association of the feature `actor` either with backward or with forward. When the length of a PIS is 3, we split it into two parts to calculate the transition probability (p_{ij}): (I) the probability of the `actor`, given that we know the backward, and (II) the probability of the `forward`, given that we know the `actor`. When the length of a PIS is 2, we use (I) for the combination “backward+actor” and (II) for “actor+forward.” Note that the number of elements of “backward” and “forward” can be greater than or equal to 1.

The graphical representation of a first-order Markov chain is a transition diagram following the transition matrix [97]. The transition diagram can be represented by a labeled directed graph whose set of vertices is E , and for which there is a directed edge from $i \in S$ to $j \in S$ with label p_{ij} , for $p_{ij} > 0$.

Let $P(a_i, a_j) = p_{ij} \in \mathbb{R}$ and $(a_i, a_j) \in E$ respectively denote the affinity score and the edge between a_i and a_j . The graph is symmetric if and only if $(a_i, a_j) = (a_j, a_i)$, $\forall (a_i, a_j), (a_j, a_i) \in E$.

4.4 Data preparation

We investigate five online discussion datasets: a set of WhatsApp group data (WGD), a set of R community data on Twitter (#rstats), freeCodeCamp Gitter Chat data (FCC), Internet Argument Corpus V2 (IAC2) and Annotated Coarse Discourse (ACD). Our datasets include debates on political, technology-related, and miscella-

4.4. Data preparation

Table 4.2 – Dataset Information.

Dataset	Obs.	Nb. actors
WhatsApp Group Data (WGD)	4K	50
Twitter Data the R Community (#rstats)	167.6K	8.8K
freeCodeCamp Gitter Chat (FCC)	5M	400K
Internet Argument Corpus V2 (IAC2)	414K	3.5K
Annotated Coarse Discourse (ACD)	101K	9K threads

neous topics. We have selected them based on their size (Table 4.2) to evaluate the method’s versatility and how it scales from small to big datasets.

FCC consists of posts extracted from 31-Dec-2014 to 9-Dec-2017 from the public chat room of Gitter⁶ which is used for an online course on programming languages and data science. We collected Twitter data from 1-Jan-2018 to 11-Sept-2018 associated with the hashtag “#rstats”, the most popular, flagship hashtag for discussions related to the R Project for Statistical Computing. This hashtag is principally used by R users and developers all over the world to facilitate information sharing across the R community on matters such as package releases, hints, discoveries, scripts, conferences, meetups, etc., as well as quick questions. The hashtag “#rstats”⁷ can to some extent be considered as the online group communication of the R community.

We extracted WGD from 1-June-2017 to 16-June-2017 from a WhatsApp group. For privacy and ethical reasons, we avoid displaying personally identifiable information in this ODD, especially names. We therefore rendered the information anonymous by using the prefix “actor_” combined with a random number that varies from one to the total number of users of the group.

IAC2 is a collection of corpora of political debate topics on online forums [1], of which we are especially interested in one, 4forums. It includes over 414K posts for over 3.5K actors. 4forums has crowdsourced annotations with a high inter-annotator agreement for stances of users in each topic and dis/agreement between users who reply to one another.

The Annotated Coarse Discourse (ACD) dataset is a large corpus of discourse

6. <https://www.kaggle.com/freecodecamp/all-posts-public-main-chatroom/>

7. <http://www.rstats.news/about.html>

CHAPTER 4. A NEW APPROACH FOR AFFINITY RELATIONSHIP DISCOVERY IN ONLINE FORUMS

annotations and relations collected by [215] from Reddit. Its goal is to allow a better understanding of online discussions at scale. It contains over 9,000 threads comprising over 101,000 comments, manually annotated. Essentially, the discourse-act annotation scheme was developed to highlight comments that include agreement, appreciation, disagreement and negative reactions.

We preprocessed the experimental datasets by removing numbers, punctuation from each token, white spaces and non-alphabetical and special characters, except the Unicode characters that correspond to emojis, and by filtering out tokens that are stopwords; and we lowercased all word tokens. It should be noted that IAC2 and ACD are already annotated. Consequently, we use their labels directly for classifying messages, whereas for the other three experimental datasets, we first label messages as described in Section 4.3.1 and then carry out the classification. The accuracy of the sentiment classifier on WGD is 88.26% for logistic regression (LR) and 73.51% for random forest (RF); on #rstats 94.18% for LR and 85.73% for RF; on IAC2 98.91% for LR and 95.05% for RF; on FCC 97.66% for LR and 92.84% for RF; and on ACD 93.42% for LR and 90.85% for RF (see Section 4.3.1). Quantitatively, we achieved better classification accuracy with the LR classifier and definitely decided to use it for classifying the sentiment of messages from the experimental datasets.

4.5 Experiments

We compare HAR-search to the state-of-the-art method, demonstrating its robustness and ability to consider minute details in diverse settings.

HAR Graph. We use the smallest dataset (WGD) to graphically illustrate the affinity relationships between community members. We apply Markov chain models to the obtained PIS to generate the transition diagram (see Figure 4.1), where the edge labels between nodes denote affinity scores and nodes represent actors. It should be noted that this graph does not display null relationships, since we have discarded all pairs of nodes for which the affinity score is zero.

The affinity may change over time depending on the way a pair of individuals maintain their relationship. As a result, there may be an increase or a decrease in the affinity score. This means that when analyzing the entire data we get general

4.5. Experiments

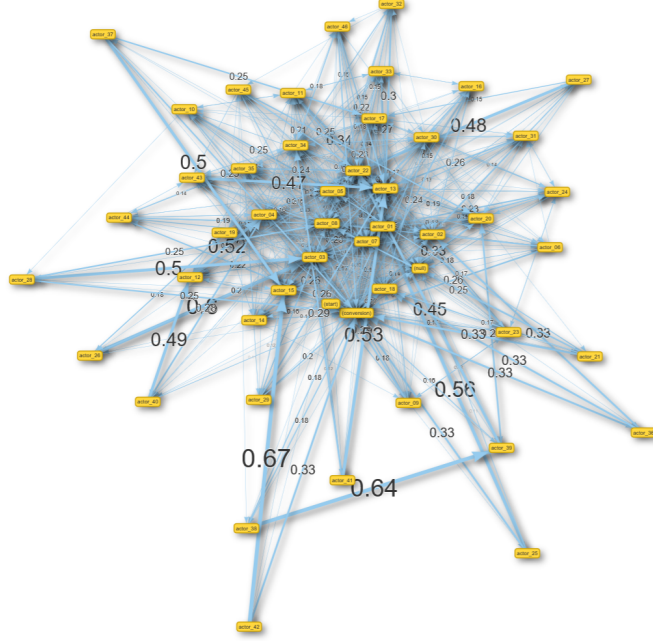


Figure 4.1 – HAR graph (positive interaction sequence-based transition diagram).

affinity scores, but when analyzing the data over time we capture affinity scores for individual periods. Based on this, we first used the whole data without any split in order to observe the overall affinity behavior, and then divided the data into time frames to learn about the affinity behavior over time.

Affinity Distribution. Using the experimental datasets (Table 4.2), we observe that the largest affinity score for the five datasets is 98.61% and the smallest is $10^{-40}\%$ (in FCC). To verify the affinity distribution, we define affinity score ranges based on our own observation, which we further group into seven segments as follows: s_1 for score $< 1\%$, s_2 for score $\in [1\%, 5\%[$, s_3 for score $\in [5\%, 10\%[$, s_4 for score $\in [10\%, 15\%[$, s_5 for score $\in [15\%, 25\%[$, s_6 for score $\in [25\%, 50\%[$ and s_7 for score $\geq 50\%$. The choice of these score ranges is arbitrary. The choice of segments depends on how one wishes to explore the affinity distribution and combine affinity relationships that have similar and dissimilar properties. The affinity distribution may vary and have a new shape if the scale and interval of the segments are modified, but this would not change the original affinity score results in any way. Figure 4.2 shows the distribution of pairs of relationships that belong to the defined score

CHAPTER 4. A NEW APPROACH FOR AFFINITY RELATIONSHIP DISCOVERY IN ONLINE FORUMS

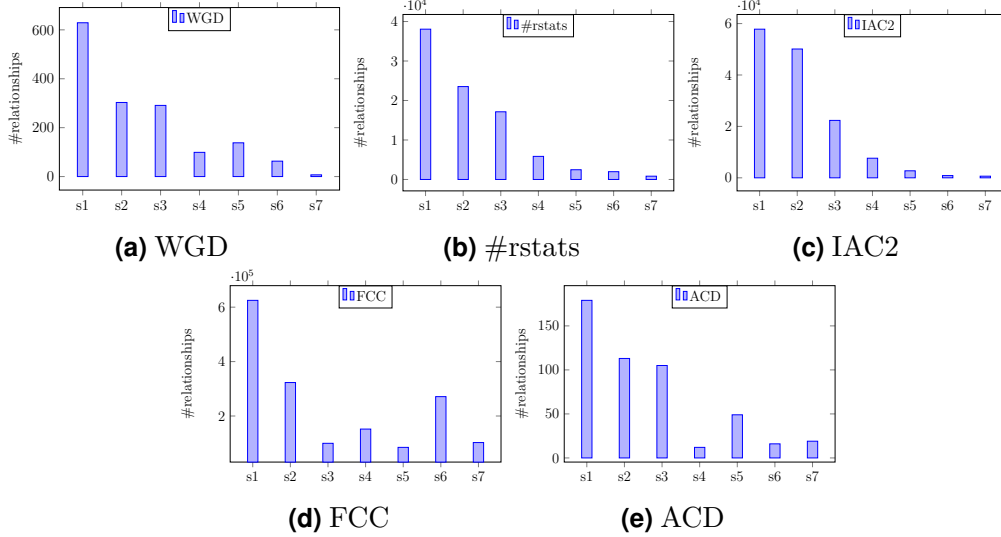


Figure 4.2 – Distribution of affinity score ranges by experimental datasets.

ranges. We found that 41%, 42%, 41%, 38% and 36% of all active relationships in WGD, #rstats, IAC2, FCC and ACD, respectively, are comprised in s_1 . Given that s_1 includes a very small affinity percentage, we assume that this segment may contain more hidden relationships than others. However, based on this assumption we pose the following question: are the hidden relationships created by uninfluential actors only, or by both uninfluential and influential actors?

Overlapped Clusters for Influence Detection. To better address this question, we formally propose definitions of influential and uninfluential actors in order to closely examine the nature of the actors in a couple of hidden relationships and provide a satisfactory response (see Definitions 1 and 2).

Definition 1 (Influential actor) *An actor a_i is influential if and only if $a_i \in \bigcap_{j=1}^m s_j$, for $j \in \{1, \dots, m\}$ and $i \in C$, where s_j denotes the segment, m is the number of segments, C represents the community and $|\bigcap_{j=1}^m s_j| > 1$.*

Definition 2 (Uninfluential actor) *An actor a_i is uninfluential if and only if a_i belongs only to s_1 ($a_i \in s_1$).*

To answer the previous question, we applied Overlapping Markov Clustering (OMC) [166] to the transition matrix of each experimental dataset. OMC is a popu-

4.5. Experiments

lar community detection algorithm to identify social groups. We obtained the social groups (or clusters) that each actor belongs to. The reason for using OMC is that an actor may simultaneously belong to many social groups, and this technique reveals all possible social groups related to her. The rationale behind this is to make it easier to verify the clusters a pair of relationships have in common.

To examine whether some actors have influence over others, we pick those actors forming the s_1 segment who do not have relationships comprised in other segments, and call them V (or uninfluential actors, see Definition 2); we select the actors forming the s_1 segment who have relationships included in other segments (regardless of the number of segments), and call them W (or influential actors, see Definition 1). For WGD, #rstarts, IAC2, FCC and ACD, respectively, we found that 2.2%, 6.1%, 5.4%, 13.7% and 9.1% of all clusters are composed solely of V. The remaining clusters involve both V and W; i.e., W play a major role in prompting uninfluential actors to become more active (Definitions 5 and 6).⁸ However, it is important to note that not all W actors are influential. Based on this hypothesis, we look for the W actors who have relationships in all segments, and call them Y, i.e., $Y \subset W$. For WGD, #rstarts, IAC2, FCC and ACD, respectively, we found that the Y and V actors together form approximately 61.1%, 23.4%, 58.1%, 16.5% and 68.3% of all clusters of hidden relationships, and the other clusters are composed of a mixture of V, W and Y actors. Based on our findings, we can say in response to the previous question that hidden relationships are created both by pairs of uninfluential actors and by pairs of uninfluential and influential actors.

Definition 3 (Terminated relationships) *We take V actors who were once in mutual relationships, but cease to have relationships over time:*

$$\{(a_i, a_k)_{t>1} \notin s_{j \geq 1} | (a_i, a_k)_{t=1} \in s_1, \forall (a_i)_{t=1} \in s_1 \text{ and } (a_k)_{t=1} \in s_1\}.$$

Definition 4 (Relationships between influential actors) *We basically explore relationships between influential actors who have and who do not have connections with V actors. Let H denote the set of influential actors (see Definition 1) and F denote the set of influential actors who have relationships with uninfluential actors (V) as described in Definition 2. We deduce K and L from these two sets as follows:*

8. This means that W actors may help V actors go beyond the boundaries by not limiting themselves exclusively to the clusters of V actors, but also opening up to W clusters.

CHAPTER 4. A NEW APPROACH FOR AFFINITY RELATIONSHIP DISCOVERY IN ONLINE FORUMS

- $K = \{H \setminus F\}$: influential actors who do not interact with V actors.
- $L = \{H \cap F\}$: influential actors who interact with V actors.

Definition 5 (Influence of W on V actors) *We investigate whether relationships between W and V actors have some impact on V actors and/or help them thrive. Essentially, we seek to demonstrate this impact in terms of evolution within the community and/or in the relationships they form with actors from other segments. We verify whether such a relationship may prompt V actors either to evolve or not to evolve. To this end, we say that V actors evolve if and only if they had relationships with W at time t , and at time $t + 1$ they no longer belong to the s_1 segment. On the other hand, we say that V actors do not evolve if and only if, after being in relationships with W at time t , they remain in the s_1 segment.*

- *Evolved*: $\{(a_i)_{t>1} \in s_{j>1} | (a_i, a_k)_{t=1} \in s_1, \forall a_i \in s_1 \text{ and } a_k \in H\}$
- *Not Evolved*: $\{(a_i)_{t>1} \in s_1 | (a_i, a_k)_{t=1} \in s_1, \forall a_i \in s_1 \text{ and } a_k \in H\}$

Definition 6 (Influence of Y on V actors) *We rigorously apply the logic used in Definition 5, only replacing W by Y , that is, the set of influential actors that belong to all segments.*

- *Evolved*: $\{(a_i)_{t>1} \in s_{j>1} | (a_i, a_k)_{t=1} \in s_1, \forall a_i \in s_1 \text{ and } a_k \in Y\}$
- *Not Evolved*: $\{(a_i)_{t>1} \in s_1 | (a_i, a_k)_{t=1} \in s_1, \forall a_i \in s_1 \text{ and } a_k \in Y\}$

Influence Detection. Beyond visualization of the affinity distribution and discovery of the influence properties of the actors, we examine the results obtained in Figure 4.2 to investigate Definitions 3, 4, 5 and 6. Specifically, Definition 3 looks for pairs of V actors who had relationships in the past, but no longer have affinity relationships over time. Definition 4 seeks to determine how influential actors maintain their relationships. To do so, we simply compare pairs of relationships of influential actors who have and who do not have relationships with V actors. Ultimately, Definitions 5 and 6 explore whether the relationships between uninfluential and influential actors are more likely to promote the evolution of uninfluential actors over time.

Terminated relationships. Based on Definition 3, we note that for WGD, #rstats, IAC2, FCC, and ACD, respectively, 35.2%, 55.6%, 18.4%, 28.9% and 17% of relationships between only V actors cease to exist over time.

4.5. Experiments

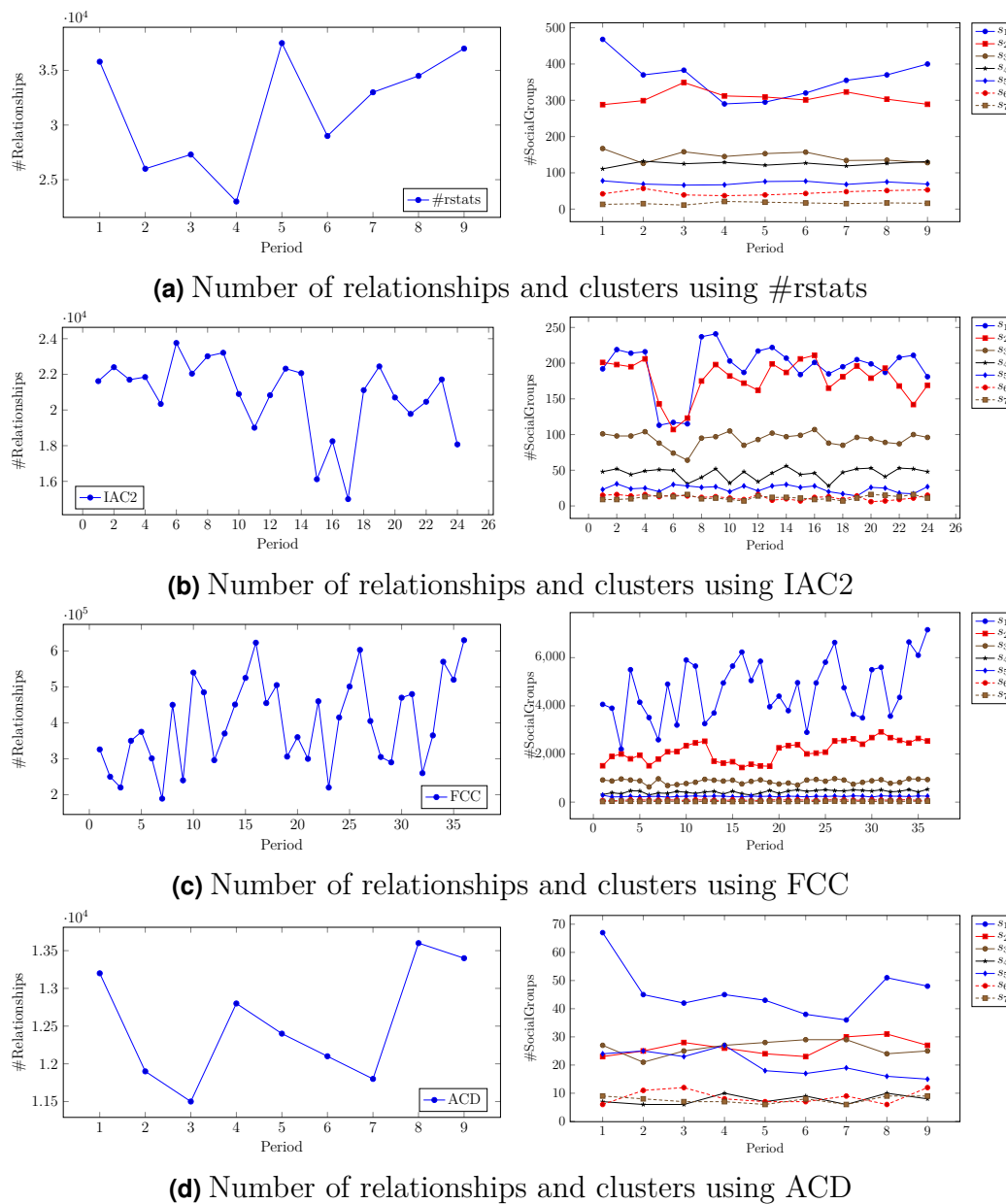


Figure 4.3 – Evolution of affinity relationships and social groups over time. Figures with `#Relationships` on the y-axis represent the count of relationships by period and show the development of relationships over time. Figures with `#socialGroups` on the y-axis give the total number of social groups by period and show their evolution over time. Note that the period is based on a one-month interval.

CHAPTER 4. A NEW APPROACH FOR AFFINITY RELATIONSHIP DISCOVERY IN ONLINE FORUMS

Relationships between influential actors. We note that there are more K actors than L actors in #rstats and FCC, and additionally observe that the K actors have developed more relationships with L actors than among themselves in all five experimental datasets (Definition 4). It should be recalled that #rstats and FCC are data science communities. It may be that K actors prefer to talk to L actors because it enables them to debate on fruitful topics likely to spark profound reflections that may be profitable to them, instead of talking to beginners and novices whose questions are less clear and concise.

Influence of W and Y over V. For WGD, #rstats, IAC2, FCC, and ACD, respectively, we report that (3.2%, 8.9%), (1.6%, 5.1%), (2.9%, 7.3%), (1.3%, 4.4%), (3.8%, 9.1%) of V actors who had relationships with W and Y actors evolved and later belonged to other segments (Definitions 5 and 6).

Affinity Evolution. To demonstrate our method’s ability to track the evolution of hidden affinities, we divide the data into periods and perform independent monthly analyses. The rationale behind this is to investigate the evolution of relationships [35] between individuals through their affinity score by examining whether it has remained constant, increased or decreased at any given time. We find that when the number of interactions rises, this yields more clusters in which the presence of individuals forming s_1 pairs is higher, and when the number of interactions drops, the number of clusters containing individuals from the s_1 pairs also decreases (see Figure 4.3).

Affinity Prediction. To make predictions over time for pairs of relationships arising through the influence of one of the actors, we consider seven features: *time*, *relation*, *affinity score*, *segment*, *affinity status*, *number of social groups (or clusters)* and *label*. The feature *relation* denotes a pair of relationships. The number of clusters is the total number of common clusters that a pair of individuals both belong to. The feature *time* implies the time frame as split above, and the value assigned to it is the concatenation of M with the number of the month. For example, FCC includes data for approximately 36 months, so the period instances vary between $M1$ and $M36$.

Specifically, we compare the affinity score for all pairs of relationships from one month to another. This allows us to assign a label of *initial* for the month when a particular relationship appears for the first time, and a label of *increase*, *decrease* or *stable* if the affinity score increases, decreases or remains constant

4.5. Experiments

based on the observed difference in the subsequent month. The feature *affinity status* indicates this change. The feature *label* indicates whether a relationship arises from any influence. Knowing that a pair of relationships basically includes two actors, if one of the actors is uninfluential and only occurs in clusters to which another actor belongs, we take this relationship as being the fruit of the influence of one actor. If both actors are uninfluential, we assume that there is no influence in the relationship. If both actors are influential, we consider that the relationship was not formed through the influence of either actor. In cases where we find influence in the relationship, we put “yes” and in other cases, we place “no”. Finally, this feature is used as the response variable to build our predictive models.

As mentioned earlier, affinity relationship estimation can be useful in many different contexts, such as recommender systems in online markets [34], political campaigns, police investigations, etc. In all of these domains, advancement requires a search for potential individuals of interest. Indeed, being able to predict the classes “yes” and “no” may tremendously contribute to the advancement of these domains in highlighting affinity relationships.

In order to demonstrate the application’s performance in generating affinity estimates adequate to provide a good solution for particular domains, some ground-truth information is required. However, ground-truth information to capture the prediction accuracy of affinity relationships is scarce and establishing it is a challenging problem. It should be noted that the lack of ground truth datasets does not affect the generalization of the findings and model performance. The results resort from the context in which the forum content has been discussed and the detection of context does not depend on annotated data. In this paper, we have opted to use variation in affinity status over time to characterize relationships in terms of their stability.

Model Performance. We used forward stepwise additive regression [111] to select the most appropriate feature set. However, we discovered that *relation* is the only feature that leads to a biased model and affects the overall performance and accuracy. We have therefore removed that feature from our list. Since we are looking at prediction over time, we prepare the training set using data from the first to the penultimate month, according to the division described above, and take the data from the last month as the test set. We then perform support vector machine (SVM),

CHAPTER 4. A NEW APPROACH FOR AFFINITY RELATIONSHIP DISCOVERY IN ONLINE FORUMS

random forest (RF) and logistic regression (LR) to measure the prediction accuracy. For SVM, we set the value γ of the radial basis function kernel to 0.5, and for RF, we built a model with 100 trees. Table 4.3 shows the performance of the models, measured by precision, recall and F-1 score. F-1 score is the harmonic mean of precision and recall and is a measure of a test’s accuracy.

Model Evaluation. To validate the performance of the proposed method, we compare it with the following baseline methods: AffinityFinder [156], WHR [173] and MIR [218]. To the best of our knowledge, AffinityFinder is the first method to address the problem of affinity detection in the context of online discussion. AffinityFinder combines structural features and content of interactions, while the two other baselines exclusively extract implicit/hidden relations based on structural features of social networks.

- AffinityFinder [156] detects affinity relationships from mentions and tweet sentiments.
- WHR: Song et al. [173] used online message threads to discover implicit relations among users who participated in the threads. Their approach considers the hierarchical relation of comments in a thread and assumes that closely related users are those for whom comments co-occur in the thread. It draws inspiration from the association rule and takes the thread as a transaction, while the frequency is utilized as a social relation for the user set. To calculate the relation score of each transaction, three different methods are suggested. We have opted to use weighted harmonic rule mining with a root-included sliding window, since it yields the best performance in terms of the harmonic mean relative to the weight of each user.
- MIR: Zhou et al. [218] proposed a social network matrix to measure the implicit relations among the entities in many social networks, including social communities. They derived several indicators to characterize the dynamics of explicit social networks along with their implicit counterparts.

To verify the effectiveness of the proposed method, we summarize the F-1 scores obtained by the models on the task of predicting the evolution of the affinity over time. We empirically show that our method outperforms the baselines in all prediction horizons by a considerable margin. Compared to AffinityFinder, HAR-search achieves

4.5. Experiments

Table 4.3 – Performance results for the proposed method and baselines on four experimental datasets. Precision, Recall and F-1 score metrics are utilized to measure model performance. Note that bold font indicates the best results for each class label per method.

Label	Method	Model	#rstats			FCC		
			Prec.	Rec.	F-1	Prec.	Rec.	F-1
Yes	HAR-search	SVM	0.859	0.817	0.837	0.867	0.845	0.856
		RF	0.873	0.856	0.864	0.881	0.852	0.866
		LR	0.862	0.848	0.855	0.873	0.866	0.869
	AffinityFinder	SVM	0.821	0.804	0.812	0.788	0.818	0.803
		RF	0.817	0.828	0.822	0.814	0.816	0.815
		LR	0.769	0.776	0.772	0.807	0.813	0.81
	MIR	SVM	0.744	0.696	0.719	0.677	0.681	0.679
	WHR	ROOT	0.661	0.662	0.661	0.645	0.652	0.648
No	HAR-search	SVM	0.898	0.797	0.844	0.832	0.847	0.839
		RF	0.864	0.753	0.805	0.857	0.872	0.864
		LR	0.875	0.871	0.873	0.833	0.856	0.844
	AffinityFinder	SVM	0.795	0.818	0.806	0.774	0.783	0.778
		RF	0.822	0.741	0.779	0.791	0.809	0.8
		LR	0.732	0.823	0.775	0.816	0.799	0.807
	MIR	SVM	0.728	0.733	0.73	0.67	0.672	0.671
	WHR	ROOT	0.685	0.687	0.686	0.639	0.64	0.639
Label	Method	Model	IAC2			ACD		
			Prec.	Rec.	F-1	Prec.	Rec.	F-1
Yes	HAR-search	SVM	0.882	0.877	0.879	0.773	0.774	0.773
		RF	0.874	0.869	0.871	0.769	0.782	0.775
		LR	0.903	0.895	0.899	0.798	0.819	0.808
	AffinityFinder	SVM	0.864	0.881	0.872	0.741	0.756	0.748
		RF	0.822	0.829	0.825	0.755	0.733	0.74
		LR	0.817	0.815	0.816	0.766	0.752	0.759
	MIR	SVM	0.656	0.658	0.657	0.67	0.668	0.669
	WHR	ROOT	0.599	0.601	0.6	0.653	0.654	0.654
No	HAR-search	SVM	0.874	0.868	0.871	0.809	0.785	0.797
		RF	0.885	0.911	0.898	0.763	0.772	0.767
		LR	0.883	0.884	0.883	0.814	0.781	0.797
	AffinityFinder	SVM	0.851	0.847	0.849	0.797	0.789	0.793
		RF	0.866	0.873	0.869	0.744	0.706	0.725
		LR	0.839	0.845	0.842	0.788	0.77	0.779
	MIR	SVM	0.662	0.656	0.659	0.664	0.671	0.667
	WHR	ROOT	0.6	0.602	0.601	0.656	0.654	0.655

better results on the experimental datasets, with an F-1 score of over 75%. Moreover, HAR-search yields statistically significant improvements over MIR and WHR.

To demonstrate the ability to predict affinity relationships, we observe that the three classifiers (SVM, RF, LR) are likely to predict both influenced and uninfluenced relationships with higher accuracy. The F-1 scores of the three models are significantly higher on the prediction of influenced relationships than for uninfluenced

CHAPTER 4. A NEW APPROACH FOR AFFINITY RELATIONSHIP DISCOVERY IN ONLINE FORUMS

relationships. We notice that LR is the best-performing model, since it provides the majority of the highest F-1 scores among the experimental results for the two methods, notably 89.9% for the proposed method on the prediction of influenced relationships on IAC2. We remark that SVM produces the second-best results in both labels for the two methods.

Further investigation reveals that the means of the F-1 scores for HAR-search over the two labels surpass those for AffinityFinder on all experimental datasets, especially with a difference of more than 5% over #rstats and FCC, and a gap of more than 2.1% on the other two experimental datasets. Specifically, we obtain the following results:

- (HAR-search = 0.852 > AffinityFinder = 0.802) for the label “yes” and (HAR-search = 0.841 > AffinityFinder = 0.787) for the label “no” on #rstats,
- (HAR-search = 0.864 > AffinityFinder = 0.809) for the label “yes” and (HAR-search = 0.849 > AffinityFinder = 0.795) for the label “no” on FCC,
- (HAR-search = 0.883 > AffinityFinder = 0.838) for the label “yes” and (HAR-search = 0.884 > AffinityFinder = 0.853) for the label “no” on IAC2,
- and (HAR-search = 0.785 > AffinityFinder = 0.749) for the label “yes” and (HAR-search = 0.787 > AffinityFinder = 0.766) for the label “no” on ACD.

We observe that the models based exclusively on structural features achieve good performance: MIR yields better results than WHR, even if both methods seem to give roughly the same results on FCC. We note that the smallest F-1 scores of AffinityFinder still surpass those of the other two methods on the four datasets. This demonstrates clear benefits and constitutes strong evidence that combining structural features and content of interactions in social networks can provide better performance on the task of affinity detection.

It should be noted that the task of affinity detection in the context of online discussion communities entails tracking user affinity degrees without necessarily taking into consideration offline inputs, such as the social, cultural, and psychological environment and socioeconomic status; or even social ties that users have offline. These opportunities remain ripe areas for future research.

4.6 Conclusion and future work

We have presented HAR-search, a method for the discovery of hidden affinity relationships between individuals within an online community. HAR-search models positive interaction sequences (PIS) based on the context of messages in the discussion history. Markov chain models are then used to quantify the PIS to yield affinity scores. These values denote the degree of affinity between a pair of community members.

In addition, the evolution of affinity over time is tracked to predict affinity relationships arising through the influence of certain individuals in the community. Finally, the results leave room for additional research. Our work provides new directions for affinity detection research in the context of online discussion communities. As future work, we plan to address both positive and negative interactions. In addition, we would like to explore the effect of negative affinity in a relationship and predict behavioral deterioration in the case of users with affinity.

Chapter 5

Discovering Affinity Relationships between Personality Types

Jean Marie Tshimula,¹ Belkacem Chikhaoui,^{1,2} Shengrui Wang¹

¹Département d'informatique, Université de Sherbrooke, QC J1K 2R1, Canada

²LICEF Research Center, Université TÉLUQ, QC H2S 3L5, Canada
{kobj2801, shengrui.wang}@usherbrooke.ca, belkacem.chikhaoui@teluq.ca

Keywords: Personality type, Affinity relationship, Emotional stability, Semantic similarity, Influence of personality on affinity, Personality prediction, Spontaneous language.

Abstract

Psychology research findings suggest that personality is related to differences in friendship characteristics and that some personality traits correlate with linguistic behavior. In this paper, we investigate the influence that personality may have on affinity formation. To this end, we derive affinity relationships from social media interactions, examine personality based on language use to discover the emotional stability of affinity relationships, and measure semantic similarity at the personality type level to understand the logic behind the development of affinity. Specifically, we conduct

5.1. Introduction

extensive experiments using a publicly available dataset containing information on individuals who self-identified with a Myers-Briggs personality type. Our results identify certain influential personality types that weigh more heavily on affinity relationships and show that personality can be predicted from the spontaneous language with an F-1 score superior to 0.76. Future research avenues are proposed.

5.1 Introduction

The study of friendship has long been a mainstay of research on developmental psychology [23, 47]. There are various stages of friendship, including formation, maintenance, and dissolution. Our focus here is on friendship formation. To some extent, the process of friendship formation can be fairly similar in real-life and online social networks, in that it involves the transition from strangers to acquaintances to friends. Individuals engage in interactions to get to know each other and forge the affective bond that characterizes a friendship. While friendships are formed differently and for various reasons, all friendships undergo a formation process. Research has shown that the formation process may be influenced by various factors, which may be environmental, individual, situational or dyadic, such as personality similarity effects [72]. Intuitively, people who share common values, tenets, convictions, and personality traits are more likely to become friends. Research on personality and friendship has yielded profound discoveries, but the two are usually studied singly; their interdependence has been investigated only recently [47, 72, 107]. This paper attempts to bridge the gap between personality and friendship by utilizing online social interactions to investigate the psychological processes underlying the development of affinity.

Our paper specifically regards affinity in friendships. To the best of our knowledge, our work is the first to address the combination of affinity and personality. Practically, investigating affinity and personality is of interest not only for psychology but also for commercial applications, including in mental health services to understand the psychological aspects and the effects of mental illness on individual patients and social systems. The combination of affinity and personality allows us to understand

CHAPTER 5. DISCOVERING AFFINITY RELATIONSHIPS BETWEEN PERSONALITY TYPES

how individuals with similar personality traits get to develop their affinity and discern what attracts an individual to another.

Most studies on personality use questionnaires (and/or written essays) to assess personal behavioral preferences. This approach inherently inhibits the expression of individual traits and makes it difficult to track language use and interactions between subjects of the survey. To efficiently conduct an analysis of language use between individuals based on their personality types requires that the data be annotated beforehand. The lack of labeled data impedes the potential of computational personality recognition to yield reliable, high-quality results [137]. It should be noted that manually annotated datasets are expensive and hard to obtain. To overcome the limitation of the small size of annotated data samples and closed-vocabulary, we chose to utilize social media data [148]. Specifically, we have collected a corpus of 758,426 English tweets with self-assessed Myers-Briggs Type Indicators [21], denoted MBTI. The MBTI assessment is based on research and personalized preferences and can contribute important information to the understanding of individual psychological functions such as intuition, sensation, thinking, feeling, etc. The MBTI model defines four binary dimensions – Introversion-Extraversion (I-E), Intuition-Sensing (N-S), Feeling-Thinking (F-T), Perception-Judgment (P-J) – that combine to yield 16 personality types into which individuals may be classified: e.g., INFP, ESTJ, ISFJ, etc. The characteristics of each MBTI personality type are described in Table 5.7.

Furnham et al. [54] performed a correlation analysis of personality traits between the MBTI and Big Five models and showed that the Big Five dimension Extraversion correlates with the MBTI (I-E), Openness to Experience correlates with (N-S), Agreeableness with (F-T), and Conscientiousness with (P-J). The rationale for using the MBTI model is that it facilitates the collection of gold-standard labeled data compared with the Big Five. The 16 MBTI personality types are simple to manipulate to account for personality differences. Since the MBTI model lacks reference to the Big Five Neuroticism dimension, we investigate the language use of individuals who self-identified with an MBTI personality type in order to discover their emotional stability. To this end, we use a psychometrically validated system to extract emotion-based psycholinguistic features. We utilize self-identified MBTI personality types as annotations and train five different models to predict personality on a linguistic level.

5.2. Related work

In order to understand the factors that contribute to the establishment of affinities, we investigate emotional stability and semantic similarity in affinity pairings based on their personality types. We seek to identify the influential personality types that weigh more heavily on affinity relationships.

To summarize, we make the following contributions:

- We show the effectiveness of our data collection and data pre-processing strategy to gather social media postings containing MBTI personality types.
- We discover personality-based affinity relationships from social media interactions and investigate the emotional stability of affinity relationships based on language use.
- We measure semantic similarity in affinity pairings at the personality level to understand the logic behind the development of affinity.
- We propose an approach to detect the influence that personality has on affinity formation.

In line with these contributions, the remainder of this paper is organized as follows. Section 5.2 discusses some related work. Section 5.3 describes the strategies utilized to extract and process our dataset. In Section 5.4, we explain our methodology, from affinity computation, through the formulation of affinity graphs with personality traits, to detect the influence of personality on affinity. We then present our experimental setup and discuss results on similarity, psycholinguistic features, and prediction in Section 5.5. In all cases, the results we obtain are thoroughly analyzed. Results are extensively discussed in Section 5.6. Finally, Section 5.7 puts forward some concluding remarks and presents future directions.

5.2 Related work

Most studies on personality and friendship rely on the most popular personality construct in contemporary psychology, the Big Five personality traits [96], to scrutinize interpersonal attraction [76, 158] and psychological well-being (satisfaction, happiness, self-acceptance, etc). Demir and Weitekamp [42] investigated the role that friendship plays in happiness and showed that friendship quality can contribute to happiness above and beyond the influence of gender and personality. Laakasuo

CHAPTER 5. DISCOVERING AFFINITY RELATIONSHIPS BETWEEN PERSONALITY TYPES

et al. [107] and Wilson et al. [200] have focused on similarities between friends and friendship patterns and found that certain personality traits are important predictors of friendship satisfaction. For instance, people who exhibit the personality traits of extroversion, agreeableness, and conscientiousness have more satisfying relationships than those who rank high in the personality trait of neuroticism. Neurotic people are linked to lower satisfaction. This may be partly explained by the fact that emotionally unstable people can be somewhat on the dramatic or high-maintenance side. Additionally, studies have shown that conscientious people have fewer unemployed friends and are more likely to have friends of the same gender, while people with high openness to experience are more likely to befriend those of different gender and ethnicity [107]. Openness to experience seems to be associated with exploratory and complementary friendship styles, while agreeableness and a lesser degree of extroversion are related to more traditional friendship ties, stressing stability and proximity of friends [107]. Extroversion, conscientiousness, and openness to experience have all been shown to influence relationship development, but their effects are inconsistent [72].

Understanding the factors that contribute to interpersonal attraction and lead to friendships can be of crucial importance. Roberts-Griffin [158] consequently focused on three factors (namely propinquity effect, similarity, and attractiveness) and found that these factors have a significant effect on whom individuals befriend. The three factors can be important when selecting close friends. Furthermore, Roberts-Griffin [158] asserted that these factors can also work in negative ways: that is, individuals can come to dislike others in the presence of these three factors.

Friendships in social media are generally inferred from structural features [9, 181]. However, relying solely on structural features may fail to extract some essential friendship character traits. For instance, in online social interactions, individuals may appear to be closer to one another based on social network structure, while they do not always show mutual appreciation and their interactions entail some divergent opinions. Tshimula et al. [186] therefore combined structural features and the content of interactions between individuals to understand their friendships and measure affinity scores between them, and predicted affinity relationships arising from the influence of certain individuals. We utilize the approach introduced by [186] to gener-

5.3. Datasets

Table 5.1 – Data summary and distribution. We collected Twitter data self-identified with MBTI personality types and calculated the percentage of each type in the dataset. We observe that INFJ comprises a large amount of data, and ISTP a much smaller amount.

Type	ISTJ	ISFP	INFP	ESFJ	ISTP	ISFJ	INFJ	ENTP	INTP	INTJ	ESFP	ENTJ	ESTP	ESTJ	ENFP	ENFJ
%	10.3	6.2	7.3	8	1.7	9.2	12.3	2.6	3.3	8.7	3.5	5.6	2.9	6.5	6.8	5.1

ate a personality-based affinity graph. We measure emotional stability and semantic similarity between affinity pairings. We then apply graph clustering to discover the connectivity between nodes within each cluster and build a methodology to detect the influence that personality has on affinity. The rationale behind the detection of the influence of personality on affinity within clusters is to identify all possible groups formed by individuals based on their interactions.

5.3 Datasets

For this research, we prepared a dataset consisting of tweets from individuals who publicly self-identified with one of the 16 MBTI personality types. Specifically, we collected tweets containing any of the 16 MBTI personality types plus the terms “MBTI”, “Briggs” and/or “Myers”. For privacy and ethical considerations, we avoid displaying personally identifiable information, especially names and pseudonyms. Consequently, we randomly replaced such information to ensure the anonymity and privacy of the data.

Dataset A. To process the data, we removed tweets written in a language other than English. We eliminated retweets and all tweets comprising more than one personality type, and removed redundant tweets. We utilized Botometer⁹, a web-based tool that uses machine learning to classify Twitter accounts as bot or human by looking at features such as friends, social network structure, temporal activity, language and sentiment. Botometer yields an overall bot score along with several other scores that provides a measure of the likelihood that the account is a bot. Bot scores display

9. <https://botometer.osome.iu.edu/>

CHAPTER 5. DISCOVERING AFFINITY RELATIONSHIPS BETWEEN PERSONALITY TYPES

on a 0-to-5 scale with zero being most human-like and five being the most bot-like. We therefore removed arbitrarily all users for which the overall bot score is higher than 2.5. We believe that accounts displaying the score of 2.5 are in the middle of the scale, and these accounts are on a relatively neutral ground. It could be difficult to classify the bot score 2.5 as human or bot. That is the reason why we consider as a bot any account displaying an overall bot score greater than or equal to 2.5. The rationale behind this is to ensure reliable data collection.

In order to thoroughly examine the language use and how it varies across each personality type, we discarded all tweets belonging to the same user in which the MBTI personality types are different. Overall, we extracted 758,426 tweets, for the same number of users. Table 5.1 outlines dataset A and shows the distribution over the MBTI personality types. We report that 9.1% of this dataset contains mentions, i.e., the @ symbol plus a username.

Dataset B. Since the algorithm of affinity relies heavily on mentions between users [186], we retrieved the most recent tweets (up to 200) for each self-identified user of dataset A. Specifically, we obtained a total of 25,253,604 tweets with an average of 33 tweets per user. We believe that in these tweets users are more likely to make use of spontaneous language in various contexts to express themselves than when they self-report or talk about their MBTI personality type in a single tweet.

Dataset for MBTI personality type prediction. The average number of words in dataset A is 27 per user, while in dataset B, there are 4,843 per user. We therefore took all tweets in dataset B for each user and labeled them with the MBTI personality type. The annotation of dataset B facilitates the extraction of behavioral patterns related to each MBTI personality type to develop a model that can predict personality on each of the 16 MBTI personality types (see Table 5.5).

5.4 Methodology

We take the set of users who publicly self-identified with an MBTI personality type (see dataset A), and verify whether relationships exist between them. To this end, we

5.4. Methodology

regard mentions in dataset B to seek to identify tweets that link these users to one another. We obtained an overall of 3,481,737 tweets bearing mentions, that is, 13.8% of the entire dataset B. We are particularly interested in tracking and investigating social mentions. The affinity algorithm, HAR-search, utilizes mentions to effectively understand their implications in social interactions, including the sentiment and the context in which mentions were tagged in discussion threads, in order to derive affinity relationships. For a good retrospective and prospective summary of the concept of affinity in social media, we refer the reader to HAR-search [186].

5.4.1 Affinity computation

Affinity relationships can basically be observed from a set of characteristics, including mutual understanding, reciprocal and common interests, sympathy, harmonious communication, and agreement between individuals [186]. In this paper, we utilize HAR-search to derive affinity scores between users from online discussions. Specifically, HAR-search considers mentions and the flow of discussions to capture minute details and contexts of interactions based on their time-series order. HAR-search extracts affinity-relevant signals from interactions, based on their sentiment and context, and then models these signals in the form of sequences. Markov chain models are then used to quantify these sequences to yield affinity scores. These values denote the degree of affinity between a pair of users. The rationale for using HAR-search is that it facilitates the generation of a Markov transition probability matrix to construct an affinity graph and track the evolution of the affinity between individuals over time, in order to predict affinity relationships arising through the influence of certain individuals within an online community. One of the added values of HAR-search is its ability to follow the temporal evolution of affinity relationships. HAR-search investigates the evolution of relationships between individuals through their affinity score by examining whether this score has remained constant, increased or decreased at any given time.

An affinity graph, $\mathcal{G} = (\mathcal{U}, \mathcal{E})$, is a weighted graph where each node $u \in \mathcal{U}$ represents a user, the edge $(u, v) \in \mathcal{E}$ denotes an affinity relation between users u and v , and the weight $w_{uv} \in \mathbb{R}$ depicts the affinity score between the two users. If edge

CHAPTER 5. DISCOVERING AFFINITY RELATIONSHIPS BETWEEN PERSONALITY TYPES

(u, v) does not exist, then the value of w_{uv} is equal to 0. In this paper, we keep only edges for which w_{uv} is greater than or equal to 10^{-5} . An affinity graph is symmetric if and only if $w_{uv} = w_{vu}$, for all $(u, v), (v, u) \in \mathcal{E}$.

Since the focus of this paper is on investigating the influence of personality on affinity formation, we refer to the affinity graph as a triplet $\mathcal{G}' = (\mathcal{U}, \mathcal{E}, \mathcal{P})$, where $\mathcal{P} = \{p_1, \dots, p_n\}$ represents the 16 MBTI personality types, \mathcal{U} is a finite node set that includes n labels, $\mathcal{E} \subseteq \mathcal{U} \times \mathcal{U}$ is an edge set and w_{uv} denotes the weight on edge (u, v) . We assign to each node u a label corresponding to an MBTI personality type, $p_i \in \mathcal{P}$. Note that each node u in \mathcal{G}' possesses only a single MBTI personality type p_i . Formally, \mathcal{L} is a mapping function for labeling nodes in \mathcal{G}' , $\mathcal{L} : \mathcal{U} \rightarrow \mathcal{P}$ such that $\mathcal{L}(u) = p_i$ is the label for node u . We assume that each node is associated with a given label in \mathcal{P} .

5.4.2 Detecting the influence of personality on affinity

To discover the influence of personality on affinity, we propose to cluster the nodes of graph \mathcal{G}' into groups of densely connected regions based on edge weights, i.e., affinity scores. To partition graph \mathcal{G}' into k overlapping clusters such that $C_1 \cup \dots \cup C_k \subseteq \mathcal{U}$, we use two different graph clustering techniques: the random walk hitting time-based digraph clustering algorithm (K-destinations) [28] and the Markov cluster algorithm (MCL) [191]. The rationale for using these algorithms in an affinity context is that they are based on first-order Markov chains, deal with directed graphs, and draw on intuition from random walks on graphs to detect cluster structure.

To determine the influence of personality on affinity, we count the number of links that each node has in a cluster C_i . Since each node is labeled with an MBTI personality type, we seek to discover nodes that include more links with nodes of different personality types within a cluster. We apply the same logic to all clusters to investigate the possible influence of personality on affinity. We assume that the overall number of links related to a specific personality type within a cluster demonstrates its openness to other types and this can be considered as a relevant signal of influence.

Here, we describe the functioning of MCL and K-destinations. MCL proposes the following intuition for the graph clustering paradigm: (i) the number of higher-length

5.5. Experiments

paths in \mathcal{G}' is large for pairs of nodes lying in the same dense cluster and small for pairs of nodes belonging to different clusters. (ii) A random walk in \mathcal{G}' that visits a dense cluster will likely not leave the cluster until many of its nodes have been visited. (iii) Considering all shortest paths between all pairs of nodes of \mathcal{G}' , edges between different dense clusters are likely to be in many shortest paths. Specifically, MCL operates on a graph where the edge weights in the graph represent similarity scores and relies on the observation that a random walk is more likely to stay in a cluster rather than travel across the clusters. MCL iteratively alternates between two successive steps of expansion and inflation until it converges. The expansion step performs random walks of higher lengths and it enables connecting to different regions in the graph. The inflation step aims to strengthen the intra-cluster connections and weaken the inter-cluster connections [162, 191].

K-destinations is an iterative clustering algorithm which uses the asymmetric pairwise measure of Markov random walk hitting time on directed graphs to cluster the data. K-destinations partitions the nodes of the directed graph into disjoint sets using the local distribution information of the data and the global structural information of the directed graph. Specifically, K-destinations suggests the following steps for graph clustering. (a) K-destinations initially fixes the destination nodes and assigns each sample to the cluster that has minimal hitting time from it to the destination node corresponding to the cluster. (b) Then, in each cluster, K-destinations updates the destination node from the samples that minimize the sum of the hitting times from all samples in the cluster to the destination node. The clustering algorithm repeats the two steps (a) and (b) until the cluster membership of each sample does not change [28].

5.5 Experiments

In this section, we show that the content of interactions between individuals self-identified with an MBTI type can be used to discover affinity relationships and analyze semantic similarity and emotional stability between affinity pairings. We further predict the MBTI personality types.

CHAPTER 5. DISCOVERING AFFINITY RELATIONSHIPS BETWEEN PERSONALITY TYPES

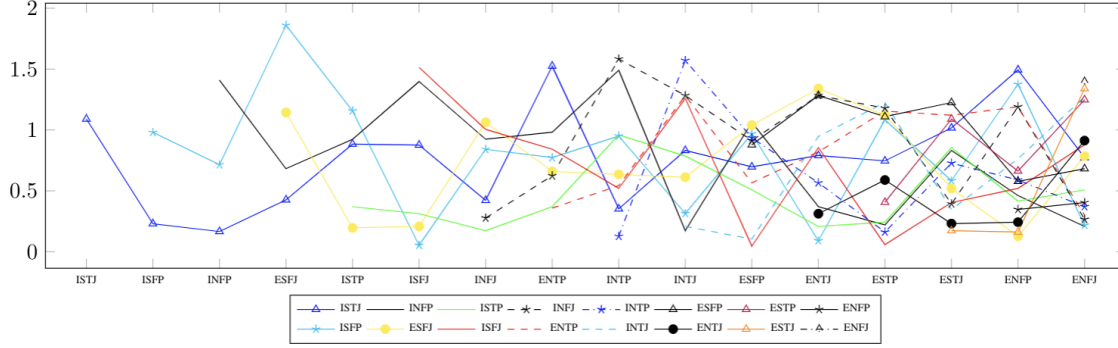


Figure 5.1 – Affinity percentages between the 136 combinations of the 16 MBTI personality types. Affinity relationships that achieve a percentage superior to 1.3% are: ENTP–ISTJ (1.53%), ENFP–ISTJ (1.49%), ESFJ–ISFP (1.86%), ENFP–ISFP (1.38%), INFP–INFP (1.41%), ISFJ–INFP (1.4%), INTP–INFP (1.49%), ISFJ–ISFJ (1.51%), INTP–INFJ (1.58%), INTJ–INTP (1.57%), ENFJ–ESTJ (1.34%) and ENFJ–ENFJ (1.4%).

Affinity discovery. In order to measure affinity relationships between individuals, we apply the HAR-search method to empirically quantify affinity connections in Dataset B. The results reported in Figure 5.1 show that affinity relationships ESFJ–ISFP and INTP–INFJ achieved the highest percentages (1.86% and 1.58%, respectively), and the affinity relationship ISFJ–ISFJ (1.51%) is the only relationship between individuals of the same personality type that reports such a high percentage. Crucially, we observe that ESFJ–ISFP and INTP–INFJ also have relatively high semantic similarity scores (Table 5.2) and low Pearson correlation coefficients for negative emotions (Tables 5.3 and 5.4).

$$\cos(D_i, D_j) = \frac{D_i D_j}{\|D_i\| \|D_j\|} = \frac{\sum_{i=1}^n a_i b_j}{\sqrt{\sum_{i=1}^n a_i^2} \sqrt{\sum_{j=1}^n b_j^2}} \quad (5.1)$$

Semantic similarity. To measure semantic similarity at the personality level, we regard only affinity relationships composed of people from two different MBTI personality types. We take all tweets belonging to people of the same personality type and assemble them in a single document (corpus). In total, we obtain 16 documents,

5.5. Experiments

Table 5.2 – Semantic similarity for affinity relationships between different MBTI personality types. Bold font indicates similarity scores greater than or equal to 0.2.

	ISTJ	ISFP	INFP	ESFJ	ISTP	ISFJ	INFJ	ENTP	INTP	INTJ	ESFP	ENTJ	ESTP	ESTJ	ENFP	ENFJ
ISTJ	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
ISFP	0.207	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
INFP	0.002	0.311	–	–	–	–	–	–	–	–	–	–	–	–	–	–
ESFJ	0.009	0.301	0.03	–	–	–	–	–	–	–	–	–	–	–	–	–
ISTP	0.214	0.208	0.005	0.01	–	–	–	–	–	–	–	–	–	–	–	–
ISFJ	0.217	0.21	0.027	0.245	0.142	–	–	–	–	–	–	–	–	–	–	–
INFJ	0.13	0.111	0.259	0.102	0.076	0.218	–	–	–	–	–	–	–	–	–	–
ENTP	0.005	0.007	0.082	0.086	0.115	0.033	0.011	–	–	–	–	–	–	–	–	–
INTP	0.08	0.1	0.304	0.002	0.243	0.007	0.094	0.232	–	–	–	–	–	–	–	–
INTJ	0.226	0.006	0.188	0.039	0.108	0.015	0.222	0.101	0.209	–	–	–	–	–	–	–
ESFP	0.04	0.209	0.076	0.227	0.062	0.11	0.003	0.095	0.082	0.003	–	–	–	–	–	–
ENTJ	0.1	0.004	0.003	0.201	0.007	0.002	0.071	0.307	0.074	0.2	0.046	–	–	–	–	–
ESTP	0.098	0.104	0.005	0.143	0.199	0.002	0.005	0.186	0.01	0.004	0.178	0.113	–	–	–	–
ESTJ	0.217	0.06	0.001	0.235	0.105	0.076	0.026	0.113	0.008	0.057	0.105	0.252	0.264	–	–	–
ENFP	0.001	0.097	0.108	0.118	0.027	0.001	0.041	0.264	0.005	0.004	0.301	0.108	0.105	0.026	–	–
ENFJ	0.002	0.006	0.101	0.253	0.001	0.104	0.06	0.119	0.013	0.025	0.114	0.257	0.061	0.114	0.289	–

$\{D_1, \dots, D_m\}$, $m = 16$. We utilize GloVe word embedding [145], an unsupervised learning algorithm for obtaining vector representations for words in each document D_i . Specifically, $D_i = \{a_1, a_2, \dots\}$ and $D_j = \{b_1, b_2, \dots\}$ denote the vector representations of two different documents, for example ISTJ and ISFP. We use cosine similarity (see Eq. (5.1)) to compute the semantic similarity of words for two documents D_i and D_j using their vector representations. The cosine similarity is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in approximately the same direction.

Table 5.2 shows semantic similarity scores for affinity pairings composed of different MBTI personality types. To understand affinity formation, we investigate the semantic similarity scores more deeply from a personality standpoint. To this end, we regard arbitrarily the threshold for affinity relationships for which similarity scores are greater than or equal to 0.2: ISFP-ISTJ (0.207), ISTP-ISTJ (0.214), ISFJ-ISTJ (0.217), INTJ-ISTJ (0.226), ESTJ-ISTJ (0.217), INFP-ISFP (0.311), ESFJ-ISFP (0.301), ISTP-ISFP (0.208), ISFJ-ISFP (0.21), ESFP-ISFP (0.209), INFJ-INFP (0.259), INTP-INFP (0.304), ISFJ-ESFJ (0.245), ESFP-ESFJ (0.227), ENTJ-ESFJ (0.201), ESTJ-ESFJ (0.235), ENFJ-ESFJ (0.253), INTP-ISTP (0.243), INFJ-ISFJ (0.218), INTJ-INFJ (0.222), INTP-ENTP (0.232), ENTJ-ENTP (0.307), ENFP-ENTP (0.264), INTJ-INTP (0.209), ENTJ-INTJ (0.2), ENFP-ESFP (0.301), ESTJ-ENTJ (0.252), ENFJ-ENTJ (0.257), ESTJ-ESTP (0.264) and ENFJ-ENFP (0.289).

CHAPTER 5. DISCOVERING AFFINITY RELATIONSHIPS BETWEEN PERSONALITY TYPES

Table 5.3 – Pearson correlations between LIWC (positive emotions) features extracted on language use to discover emotional stability in affinities between two different personality types. All correlations are significant at $p < 0.01$.

	ISTJ	ISFP	INFP	ESFJ	ISTP	ISFJ	INFJ	ENTP	INTP	INTJ	ESFP	ENTJ	ESTP	ESTJ	ENFP	ENFJ
ISTJ	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
ISFP	0.041	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
INFP	0.222	0.046	–	–	–	–	–	–	–	–	–	–	–	–	–	–
ESFJ	0.065	0.272	0.054	–	–	–	–	–	–	–	–	–	–	–	–	–
ISTP	0.249	0.114	0.041	0.107	–	–	–	–	–	–	–	–	–	–	–	–
ISFJ	0.207	0.201	0.103	0.214	0.276	–	–	–	–	–	–	–	–	–	–	–
INFJ	0.113	0.011	0.070	0.106	0.204	0.212	–	–	–	–	–	–	–	–	–	–
ENTP	0.302	0.300	0.296	0.219	0.315	0.283	0.270	–	–	–	–	–	–	–	–	–
INTP	0.260	0.285	0.248	0.315	0.297	0.250	0.268	0.313	–	–	–	–	–	–	–	–
INTJ	0.318	0.209	0.287	0.107	0.288	0.265	0.313	0.285	0.322	–	–	–	–	–	–	–
ESFP	0.066	0.058	0.032	0.043	0.174	0.109	0.047	0.041	0.041	0.197	–	–	–	–	–	–
ENTJ	0.134	0.201	0.079	0.176	0.256	0.314	0.311	0.309	0.295	0.320	0.102	–	–	–	–	–
ESTP	0.217	0.123	0.009	0.150	0.162	0.091	0.079	0.252	0.173	0.088	0.076	0.222	–	–	–	–
ESTJ	0.311	0.308	0.325	0.238	0.249	0.248	0.254	0.238	0.038	0.024	0.104	0.274	0.215	–	–	–
ENFP	0.036	0.106	0.022	0.129	0.151	0.056	0.116	0.214	0.106	0.017	0.028	0.127	0.109	0.212	–	–
ENFJ	0.205	0.223	0.087	0.183	0.295	0.237	0.202	0.207	0.209	0.195	0.170	0.251	0.203	0.194	0.083	–

Based on the preceding, it can be seen that ENFJ, ENFP, INFP, ENTJ, ENTP, ESFJ, ESTJ, and INTP each appear in two or three affinity relationships for which the similarity scores are superior to 0.23. Specifically, the types ESFJ, ENTJ, and ENFP have the highest semantic similarity scores with ENFJ. Moreover, we note a number of affinity relationships with low semantic similarity scores: INFP-ISTJ (0.002), ENFP-ISTJ (0.001), ENFJ-ISTJ (0.002), ESTJ-INFP (0.001), INTP-ESFJ (0.002), ENFJ-ISTP (0.001), ENFP-ISFJ (0.001), ESFP-INFJ (0.003) and ESFP-INTJ (0.003).

Emotional stability. To measure emotional stability in affinity relationships between two different personality types, we investigate language use in their discussion interactions and utilize the Linguistic Inquiry and Word Count (LIWC) text-analysis system to extract psycholinguistic features. LIWC is a widely used, psychometrically validated system for psychology-related language analysis and word classification. The LIWC dictionary includes word categories that have pre-labeled meanings created by psychologists. The LIWC categories have also been independently evaluated for their correlation with psychological concepts [144]. For each tweet, we computed the number of observed words and terms using the LIWC system and focusing exclusively on two LIWC categories: psychological processes and linguistic dimensions. For the psychological processes, we focused especially on the following two subcategories: positive and negative emotions. With regard to the linguistic dimensions

5.5. Experiments

Table 5.4 – Pearson correlations between LIWC (negative emotions) features extracted on language use to discover emotional stability in affinities between two different personality types. All correlations are significant at $p < 0.01$.

	ISTJ	ISFP	INFP	ESFJ	ISTP	ISFJ	INFJ	ENTP	INTP	INTJ	ESFP	ENTJ	ESTP	ESTJ	ENFP	ENFJ
ISTJ	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
ISFP	0.093	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
INFP	0.001	0.215	–	–	–	–	–	–	–	–	–	–	–	–	–	–
ESFJ	0.007	0.003	0.086	–	–	–	–	–	–	–	–	–	–	–	–	–
ISTP	0.002	0.107	0.013	0.022	–	–	–	–	–	–	–	–	–	–	–	–
ISFJ	0.001	0.083	0.025	0.008	0.021	–	–	–	–	–	–	–	–	–	–	–
INFJ	0.085	0.236	0.114	0.013	0.017	0.004	–	–	–	–	–	–	–	–	–	–
ENTP	0.002	0.003	0.001	0.007	0.001	0.001	0.002	–	–	–	–	–	–	–	–	–
INTP	0.001	0.002	0.003	0.002	0.001	0.002	0.003	0.001	–	–	–	–	–	–	–	–
INTJ	0.001	0.007	0.002	0.003	0.002	0.001	0.001	0.002	0.001	–	–	–	–	–	–	–
ESFP	0.074	0.119	0.068	0.209	0.019	0.026	0.034	0.037	0.075	0.013	–	–	–	–	–	–
ENTJ	0.001	0.002	0.002	0.003	0.001	0.002	0.001	0.001	0.002	0.001	0.009	–	–	–	–	–
ESTP	0.010	0.048	0.207	0.027	0.034	0.077	0.018	0.004	0.008	0.013	0.025	0.008	–	–	–	–
ESTJ	0.001	0.001	0.003	0.009	0.005	0.003	0.004	0.016	0.004	0.036	0.031	0.004	0.009	–	–	–
ENFP	0.206	0.004	0.116	0.118	0.023	0.115	0.015	0.021	0.013	0.034	0.116	0.012	0.016	0.007	–	–
ENFJ	0.003	0.002	0.019	0.015	0.004	0.008	0.006	0.013	0.007	0.008	0.083	0.001	0.205	0.009	0.011	–

category, we measured solely the proportion of first-person pronouns in the tweet content. Research shows that pronouns reveal information on a person’s emotional state, thinking, and personality [144]. Chung and Pennebaker [37] found that individuals who are strongly susceptible to emotional reactions or vulnerable situations more frequently use first-person pronouns, suggesting higher self-attention focus.

Pearson correlations from LIWC features. We performed linear regression with elastic-net regularization to calculate Pearson correlation coefficients, using the weights of the LIWC features. Let $X = \{x_1, x_2, \dots\}$ and $Y = \{y_1, y_2, \dots\}$ denote two feature vectors extracted from the language use of two different personality types. To compute Pearson’s r , we took the top n elements of each vector in descending order ($n=1000$); the complete results can be seen in Tables 5.3 and 5.4.

Tieger and Barron-Tieger [183] explored the personality type of many couples and found that the more type preferences a couple had in common, the more satisfied they were with their communication. In the work reported here, we found that the personality types bearing the preferences S (sensing) and J (judgment), that is, ESTJ, ESFJ, ISTJ and ISFJ, are not emotional when they are in affinity relationships among themselves. We also found that ENTP, INTP, INTJ, ENTJ and ESTJ maintain good affinity relationships with all personality types and tend to be emotionally

CHAPTER 5. DISCOVERING AFFINITY RELATIONSHIPS BETWEEN PERSONALITY TYPES

Table 5.5 – Prediction results of MBTI personality types. Bold font indicates the best performance for each MBTI type.

	LR	RF	SVM	NB	BERT
ISTJ	0.753	0.782	0.786	0.721	0.891
ISFP	0.761	0.777	0.766	0.711	0.773
INFP	0.774	0.802	0.774	0.698	0.800
ESFJ	0.780	0.783	0.776	0.763	0.785
ISTP	0.782	0.769	0.782	0.735	0.888
ISFJ	0.755	0.757	0.739	0.697	0.812
INFJ	0.775	0.768	0.757	0.747	0.774
ENTP	0.779	0.773	0.782	0.709	0.781
INTP	0.754	0.756	0.745	0.705	0.859
INTJ	0.798	0.785	0.795	0.722	0.803
ESFP	0.759	0.760	0.757	0.698	0.866
ENTJ	0.781	0.778	0.769	0.767	0.887
ESTP	0.758	0.757	0.763	0.760	0.762
ESTJ	0.784	0.789	0.773	0.755	0.894
ENFP	0.780	0.766	0.782	0.699	0.806
ENFJ	0.767	0.782	0.774	0.731	0.861

stable people (Tables 5.3 and 5.4). Our results support the outcomes of the study conducted by [183] on couples and personality type, except for ENTP, INTP, INTJ, ENTJ and ESTJ. Tieger and Barron-Tieger’s research found that (i) ESTJ, ESFJ, ISTJ and ISFJ have a satisfaction rate of 79% when paired with each other, and (ii) ENFP, INFP, ENFJ and INFJ have a satisfaction rate of 73% when paired with each other. These tend to place a high value on relationships and are the most likely of all the types to devote themselves to healthy relationships and open communication.

MBTI prediction. In order to predict each of the 16 MBTI personality types, we trained five different classifiers: logistic regression (LR, 10^8 ridge), random forest (RF) with AdaBoost, support vector machine (SVM), a simple naive Bayes (NB) and BERT [43]. For SVM, we set the regularization parameter λ to 0.0001 and the value γ of the radial basis function kernel to 0.5; for RF, we set the number of trees to 500 and the maximum depth and number of features to 3 and 30, respectively. For BERT, we used the BERT-large-cased model, which comprises 24 layers, 16 attention heads and 340 million parameters. We conducted multi-class classification by extracting and analyzing linguistic patterns from the user tweets and personality labels mentioned in Dataset B (see Section 5.3).

To evaluate the performance of the constructed multi-class classifiers, we per-

5.5. Experiments

Table 5.6 – Clustering results in terms of Error and NMI. Bold font indicates the best performances.

	MCL	5-destinations	10-destinations	15-destinations
NMI	0.822	0.848	0.814	0.797
Error	0.016	0.013	0.030	0.071

formed 10-fold cross-validation to split our training and testing sets and computed the F-1 score metric to measure the accuracy of our classifiers. Table 5.5 presents the performance results of the five classifiers. We report that the F-1 scores for our classifiers are relatively high and show the ability to predict all of the 16 MBTI personality types. It can be observed that the majority of the best performances were achieved by BERT, with F-1 scores of over 0.8. Even BERT’s poorer results outperforms some of the other classifiers by a significant margin. Interestingly, we found that ESTJ, ENTJ and ISTP were easily predicted by the five classifiers utilized, as they yielded the highest average performances: 0.799, 0.796 and 0.791, respectively. In particular, it can be seen that the personality types containing the preferences T (thinking) and J (judgment) yielded higher average performance. We also note that RF consistently performed well on the personality types bearing the preferences I (introversion), F (feeling) and P (perception), and SVM surpassed all classifiers on the personality types that include the preferences E (extraversion), T (thinking), and P (perception).

Influence of personality on affinity. To discover the influence that the personality types have on affinity formation, we utilized the approach proposed in Section 5.4.2. MCL is an unsupervised graph clustering algorithm. For K-destinations, we set the number of destination nodes by varying K, assigning it values of 5, 10 and 15. This variation allows us to better explore the influential personality types on various facets.

To evaluate the performances of MCL and K-destinations, we computed two performance measures from the clustering results: the normalized mutual information (NMI) and the minimal clustering error (Error). The NMI is defined as

$$\text{NMI} = \frac{I(x, y)}{\sqrt{H(x)H(y)}}, \quad (5.2)$$

CHAPTER 5. DISCOVERING AFFINITY RELATIONSHIPS BETWEEN PERSONALITY TYPES

where $I(x, y)$ is the mutual information between the true x and y , and $H(x)$ and $H(y)$ are the entropies of x and y , respectively. Note that $0 \leq \text{NMI}(x, y) \leq 1$ and $\text{NMI}(x, y) = 1$ when $x = y$. The larger the value of NMI, the better the clustering result.

The clustering error is defined as the minimal classification error among all possible permutation mappings, defined as:

$$\text{Error} = \min(1 - \frac{1}{n} \sum_{i=1}^n \delta(y_i - \text{perm}(c_i))), \quad (5.3)$$

where y_i and c_i are the true class label and the obtained clustering result of x_i , respectively, and $\delta(x, y)$ is the delta function that equals 1 if $x = y$ and 0 otherwise.

The clustering results for the two methods are summarized in Table 5.6. Our results demonstrate that we achieved good performance for graph clustering. It can be seen that 5-destinations achieved significantly better performance on both evaluation metrics, that is, the smallest error and the largest NMI values. Moreover, we observe that the error values for K-destinations become significantly larger as the set value of K increases, showing that this variation can reduce the NMI value by a considerable margin. Specifically, we extracted 6, 4, 7 and 9 clusters with MCL, 5-destinations, 10-destinations and 15-destinations, respectively. As described in Section 5.4.2, we counted the number of links that each node has in each cluster. We assume that the total number of links in a set of nodes indicates a relevant signal of influence. Specifically, for each cluster, we report only the node with the highest number of links. We obtained (ENTJ, ENFP, ESTJ, INTP, ISTJ, INFP) for MCL, (ESTJ, INTP, ENFP, ENTJ) for 5-destinations, (INFP, ENFP, ISTJ, INTJ, ESTJ, ENTJ, INTP) for 10-destinations and (INFP, INTJ, INTP, ISTP, ENFP, ESTP, ENTJ, ISTJ, ESTJ) for 15-destinations. Note that the four influential personality types extracted from 5-destinations are also part of MCL, 10- and 15-destinations.

5.6 Discussion

Our results provide some of the first insights into the investigation and understanding of affinity relationships between personality types on social media. We

5.6. Discussion

measured semantic similarity and emotional stability in affinities, and showed the feasibility of applying clustering to discover the influence of personality on affinity. Moreover, we trained five different classifiers from the spontaneous language utilized by a set of social media users to predict the 16 MBTI personality types. The theoretical and practical implications of our outcomes can be valuable for supporting decision-making processes in various domains, including clinical psychology, forensic psychology, digital forensics, human factors and social science.

Our results identify a number of statistically significant correlations in terms of emotional stability in personality-based affinity relationships. It should be recalled that our investigation was limited to extracting LIWC features and measuring correlation coefficients related to emotional stability in affinity pairings. This study does not specifically examine the reasons or the circumstances in which emotional reactions were expressed. Importantly, we report 13 affinity pairings for which correlation values for negative emotions surpassed 0.1: ENFP-ISTJ (0.206), INFP-ISFP (0.215), ISTP-ISFP (0.107), INFJ-ISFP (0.236), ESFP-ISFP (0.119), INFJ-INFP (0.114), ESTP-INFP (0.207), ENFP-INFP (0.116), ESFP-ESFJ (0.209), ENFP-ESFJ (0.118), ENFP-ISFJ (0.115), ENFP-ESFP (0.116) and ENFJ-ESTP (0.205). We note that only ENFP and ISFP appear in five and four different affinity pairings, respectively. Our findings show strong evidence that the types ENFP and ISFP are particularly emotionally reactive and predominantly mention negative emotions in their narratives. The two types appear quite close in terms of affinity percentage (1.38%, see Figure 5.1) and have in common two preferences (F and P). Moreover, we note that the 13 aforementioned affinity pairings have relatively high semantic similarity scores, except for ENFP-ISTJ, ESTP-INFP, ENFP-ISFJ and ENFJ-ESTP. From our experiments, our observation is that emotional stability does not depend strongly on semantic similarity. For instance, we find that affinity pairings with semantic similarity scores less than or equal to 0.003 have high and low correlation values for positive and negative emotions, respectively, except for ENFP-ISTJ, ENFP-ISFJ and ESFP-INFJ. We believe that the findings on semantic similarity and emotional stability constitute an important lead for understanding the implications of personality in the development of affinity.

An interesting thing to note about the cluster analysis is that our findings suggest

CHAPTER 5. DISCOVERING AFFINITY RELATIONSHIPS BETWEEN PERSONALITY TYPES

the value of K can greatly affect the ability of K -destinations to accurately detect clusters in the affinity graph. We therefore explored the clusters detected by both MCL and K -destinations to extract influential personality types. Before proceeding further, it should be noted that we limited ourselves to identifying the influence of personality on affinity. Applying our approach to the results yielded by the clustering techniques used, we identify potential influential personality types for each cluster and observe that these influential personality types overlap from one technique to another. We remark that the four influential personality types stemming from 5-destinations can also be found in MCL, 10-destinations and 15-destinations. However, analyzing the aspects on which certain personality types were influenced by the influential MBTI types yielded by 5-destinations constitutes a challenging problem and naturally requires further inquiry. As mentioned earlier, most studies on personality rely more heavily on questionnaires to evaluate individual preferences and predict team dynamics [126]. Combining social influence-based behavior questionnaires and social media interactions may possibly reveal important factors that can help investigate and explain the causes of influence with sufficient certainty. In reality, investigations into the influence of personality can be driven by the concrete needs of applications. Examples might be investigating the role that personality plays in the effective functioning of behavioral deterioration [187]. Our results also contribute to understand affinity-seeking behaviors and affinity-maintaining patterns between relationships of individuals of different personality types. Our approach can be used as baseline to detect affinity-seeking behaviors from textual data stemming from social media.

Applying social media users' self-identified types and examining their spontaneous language, we extracted linguistic patterns using five different classifiers to predict the 16 MBTI personality types. The results are very encouraging and show that our classifiers can effectively predict personality with high accuracy. In particular, we achieved the majority of the best performances with BERT. BERT predicted the personality by not only considering self-reported type classes but also capturing the context in which the text corpus related to each type class was expressed. To validate the performance of the classifiers used, we considered self-identified types as ground truths. A major advantage of using the self-identified types as ground truths is their ability to act as immediate validation [53]. We recognize that an individ-

5.7. Conclusion

ual’s personality could possibly develop and change over time [16]. To predict the personality types, we consistently pre-processed the experimental dataset to remove individuals who have reported two or more types. In the future, we would like to keep individuals with several self-identified types, in order to investigate the dynamic nature of personality. We believe that data processing can potentially contribute to the ever-challenging task of personality prediction from social media text data.

5.7 Conclusion

In this paper, we presented a series of analyses to understand affinity relationships between personality types on social media. Specifically, we focused strongly on individuals who self-identified with one of the MBTI types, and explicitly tracked their language use. Our results have shown significant correlations in emotional stability in affinity relationships between individuals from different personality types, and examined the semantic similarity in these affinity relationships. In addition to these analyses, we have provided new insights for discovering the influence that certain personality types have on others and predicting personality by utilizing the linguistic patterns extracted directly from spontaneous language. Our study contributes to the body of research on personality, with a new understanding of the implications of the influence that personality has on affinity relationships.

While the scope of our study is limited to understanding the influence of personality on affinity by utilizing psycholinguistic features, our findings point the way for future investigations of broader scope. For instance, in exploring the influence of personality on affinity, the socioeconomic status and demographic information of individuals could be considered. We believe that this may provide additional insights, allowing examination of more subtle details that could help to better explain the influence of personality. Future studies may juxtapose psycholinguistic and demographic features to explore different facets of the influence of personality. Moreover, we aim to utilize demographic features to measure the correlation between socioeconomic status and affinity relationships.

CHAPTER 5. DISCOVERING AFFINITY RELATIONSHIPS BETWEEN PERSONALITY TYPES

Table 5.7 – Characteristics of the MBTI types. Source: The Myers-Briggs Company (<https://eu.themyersbriggs.com/>)

MBTI	Characteristics
ISTJ	People with ISTJ preferences are typically thorough, conscientious, realistic but also systematic and reserved.
ISFP	ISFPs are cooperative, modest and adaptable and also gentle and loyal.
INFP	INFPs are flexible, spontaneous as well as reflective and contained. They are also imaginative and developmental.
ESFJ	ESFJs are warm and appreciative as well as organized, outgoing and supportive. They are also realistic and loyal.
ISTP	ISTPs are analytical, practical, realistic but also logical and adaptable.
ISFJ	ISFJs are organized, practical and patient, but also dependable and loyal. Furthermore ISFJs are patient and understanding.
INFJ	INFJs are compassionate, idealistic as well as imaginative and visionary. They are also sensitive and reserved.
ENTP	ENTPs are emergent, theoretical and flexible as well as imaginative and challenging.
INTP	INTPs are independent and detached, who also tend to be challenging and logical as well as skeptical and innovative.
INTJ	INTJs are strategic and conceptual as well as innovative, independent and logical. They can also be demanding and reflective.
ESFP	ESFPs are tolerant and spontaneous as well as playful and resourceful. They also tend to be friendly and enthusiastic.
ENTJ	ENTJs are structured and challenging, they also tend to be strategic and questioning.
ESTP	ESTPs are analytical, outgoing and enthusiastic as well as logical and they tend to be observant and resourceful.
ESTJ	ESTJs are responsible and efficient but they can also be assertive as well as logical and realistic.
ENFP	ENFPs are friendly and expressive as well as innovative and energetic.
ENFJ	ENFJs are warm, collaborative and supportive, as well as friendly and organized. They also tend to be persuasive.

Part III

Stance Detection and Behavioral Deterioration in Discussions

Summary

In this part, we investigate the problem of behavioral deterioration and propose new models that construct behavioral sequences from temporal behaviors exhibited by individuals. Since this problem is relatively new in the context of social media, we study how the divergence of opinion can potentially lead to unhealthy conversations and emotional reactions, and introduce a formal definition of the problem of behavioral deterioration. For stance classification, we construct a model on top of RoBERTa to classify stances by capturing the context of the discussion through the examination of pairs of stances and relational structures of discussion specific to each topic within the defined window of interactions of each participant of the discussion. We investigate the degree of disagreement and neutrality in the discussion to measure the divergence of opinion on topics addressed in the discussion, and predict the emotion associated with interactions by topic. For the prediction of behavioral deterioration, we propose new models to extract consecutive combinations of sequential patterns corresponding to misbehavior to predict behavioral deterioration at the individual level. We find that relying solely on individual level features to predict deterioration, in of itself, is not necessarily problematic, but this may render a significant proportion of deterioration patterns an untapped resource of potential. Consequently, we investigate temporal deterioration patterns from behavioral sequences to predict deterioration at the community level. Our experiments suggest that our models have the potential of leveraging behavioral sequences for predicting signals relevant to deterioration from accumulations of behaviors and show the ability of our models in predicting behavioral deterioration with a high degree of accuracy, i.e., F-1 scores of over 0.8. Furthermore, we examine the trajectory of behavioral deterioration in order to discover the emotional states that individuals gradually exhibit and evaluate whether these emotional states lead to the deterioration of behaviors as time moves forward. Our results suggest that *anger* could be a potential emotional state that can substantially contribute to behavioral deterioration.

Publications

Chapters 6 and 7 have been published as conference papers. Specifically, Chapter 6 is reprinted with permission from “A Pre-training Approach for Stance Classification in Online Forums” by Jean Marie Tshimula, Belkacem Chikhaoui, and Shengrui Wang. In: Proceedings of 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 280–287. Chapter 7 is reprinted with permission from “On Predicting Behavioral Deterioration in Online Discussion Forums” by Jean Marie Tshimula, Belkacem Chikhaoui, and Shengrui Wang. In: Proceedings of 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 190–195. Chapter 8 has been submitted for publication.

In these papers, Jean Marie Tshimula contributed as the first author. Jean Marie Tshimula designed the proposed model, conducted the experiments and wrote the papers. The drafting and verification of the equations were all accompanied by advisors Dr. Shengrui Wang and Dr. Belkacem Chikhaoui.

Chapter 6

A Pre-training Approach for Stance Classification in Online Forums

Jean Marie Tshimula,¹ Belkacem Chikhaoui,^{1,2} Shengrui Wang¹

¹Département d'informatique, Université de Sherbrooke, QC J1K 2R1, Canada

²LICEF Research Center, Université TÉLUQ, QC H2S 3L5, Canada
{kobj2801, shengrui.wang}@usherbrooke.ca, belkacem.chikhaoui@teluq.ca

Keywords: Opinion, Sentence-pair, Divergence of opinion.

Abstract

Stance detection is the task of automatically determining whether the author of a piece of text is in favor of, against, or neutral towards a target such as a topic, entity, or claim. In this paper we propose a method based on RoBERTa to classify stances by capturing the context of the discussion through the examination of pairs of stances and relational structures of debates specific to each topic within the defined window of each forum participant's interventions. Furthermore, we examine the degree of disagreement and neutrality in various debate topics to measure divergence of opin-

6.1. Introduction

ion in the course of the debate and estimate the emotional state manifested in different debate topics. We conduct extensive experiments using two publicly available datasets and demonstrate that our method considers more stance classes, provides better results and yields statistical improvements over existing techniques. Our quantitative analysis of model performance yields F-1 scores of over 0.745. Interestingly, we obtained the highest F-1 score, 0.814, on a stance class which was not taken into consideration in prior work. We report that none of the metrics utilized to measure divergence of opinion yield values exceeding 50% and the correlations between the same topics over 10-fold cross-validation are statistically significant for the majority of them ($p < 0.005$). Several future research avenues are proposed.

6.1 Introduction

Online forums are Internet-based group communities that provide an environment in which numerous topics can be discussed with other people who may be like-minded or hold opposing views. Since online discussion forums may assemble participants who have diversified convictions and beliefs on the various matters covered in the community, there can be a strong divergence of opinion in debates and some difficulty in reaching complete unanimity on a debate topic.

Online debates may contain discussions on several topics involving many participants, in which each intervention is either a response to a preceding post or the root of the discussion. The work reported here focuses on classifying the stances expressed by forum participants. Stance classification can be considered as the task of inferring from the text whether a particular forum participant agrees with an opinion expressed by another participant, disagrees with it, or has a neutral point of view towards it [48, 102, 117, 190]. Early work [7, 172] considered this issue as a binary classification task and focused on feature representations. It demonstrated that stance classification in the context of online forums is a very challenging problem.

Understanding stance can be of practical interest to many stakeholders, including

CHAPTER 6. A PRE-TRAINING APPROACH FOR STANCE CLASSIFICATION IN ONLINE FORUMS

companies and governments, since it can provide critical insight into the theoretical foundation of discourse, argumentation, and sentiment [49, 174]. Such knowledge can be used for multiple purposes, such as predicting behavioral deterioration, detecting affinity relationships [185, 186], revealing misinformation [146], identifying fickle-minded people and weathervanes, recognizing logical fallacies like strawman arguments, targeting public awareness and advocacy campaigns [175], adapting users' information preferences to their beliefs and ideologies, conducting personality tests [147] and online background checks, discerning the divergence of online discussion [152], and so on.

Prior studies proposed various techniques for detecting and classifying stance in a set of real-world texts. For instance, Sridhar et al. [175] introduced a collective classification method that captures the debate structure tree by modeling the dependencies between forum participants and their posts. Li et al. [117] used the structural dependencies of debate dialogues by measuring the similarity between embedding representations of a post and a given stance label. To determine the stance label, Sridhar et al.'s approach exploits manually written predicates and probabilistic soft logic to model reply links, and Li et al.'s approach relies exclusively on inference over the relationships between the learned representations of a post of interest; while other approaches merely detect the stance of participants from analysis of the text of a single post [29, 48, 49, 102, 117, 150, 178, 190, 212].

To overcome the limitations of the research discussed above, we propose a method that extracts the context of the discussion. The rationale behind context extraction is to capture relational structures of the discussion specific to each topic in order to classify proper pairs of posts. To classify two posts, we use RoBERTa [122], a Transformer-based pre-trained language model that carefully tunes hyperparameters and training data size, leading to significantly improved results on language understanding. It should be noted that RoBERTa is one of the top pre-trained language models and yielded state-of-the-art results on many NLP downstream tasks and benchmarks (see SuperGLUE¹⁰), including sentence-pair classification. We use RoBERTa to generate features and then heuristically map entailment-type class labels onto an Agreement-Disagreement-neutral relational structure to train

10. <https://super.gluebenchmark.com/leaderboard>

6.2. Related work

secondary classifiers. Furthermore, we investigate the degree of disagreement and neutrality in different debate topics to measure divergence of opinion within the debate. Specifically, this work makes the following contributions:

- We propose a method that extracts the context of the discussion by exploring possible combinations of pairs of posts specific to each topic within the window between previous and next opinions expressed by each forum participant.
- We suggest a new transformation of the discussion to capture relational structures of the debate and simplify it to build our classifiers by means of RoBERTa-based sentence-pair labels, topics and new features extracted from initial annotations of the experimental datasets.
- We explore topic-based graphs to measure divergence of opinion throughout the discussion.
- We conduct extensive experimentation using real datasets to validate the stance classification, measure divergence of opinion within the debate and assess the emotional state manifested in different debate topics.

The rest of this paper is organized as follows. In Section 6.2, we briefly outline some related work. Section 6.3 describes the proposed method. We conduct experiments and discuss the outcomes in Section 6.4. Finally, we conclude and present future directions in Section 6.7.

6.2 Related work

Prior work on stance classification has focused on linguistic features for identifying clues from oppositional speakers [7, 172], structural features for modeling agreement or disagreement with forum posts by inferring their labels [117, 174, 175, 195], sentiment [102] and neural attention network approaches [29, 48, 178, 212].

To classify stances in tweets, Tutek et al. [190] designed both lexical and task-specific features to train and fine-tune several classifiers using a genetic algorithm. Ebrahimi et al. [49] proposed a probabilistic approach that discriminates sentiment- and target-specific features and then regularizes this on a single classifier. Krejzl and Steinberger [102] constructed a maximum entropy classifier based on surface-level, sentiment, and domain-specific features. The limitation of this work is that it

CHAPTER 6. A PRE-TRAINING APPROACH FOR STANCE CLASSIFICATION IN ONLINE FORUMS

classifies stances without comparing them to one another to capture the context in which these stances were expressed.

More recently, neural attention network techniques have been applied to classify stance more efficiently. These have achieved competitive results and helped stance classification make an important stride forward to investigate new avenues. For instance, Sun et al. [178] proposed a hierarchical attention network to weigh the importance of various linguistic information and learn the mutual attention between the document and the linguistic information. Zarrella and Marsh [212] suggested a transfer learning-based method using large unlabeled datasets to learn sentence representations. Du et al. [48] introduced a neural attention model to extract target-specific related information for classifying stance in texts. These efforts were handicapped because they relied heavily on comparing post content with annotated labels, rather than classifying from the context of the stance of the hypothesis with respect to the premise.

We drew our inspiration from the work cited above, including [117] and [175] discussed in the previous section. These authors used probabilistic soft logic to model post stance by leveraging both the local linguistic features and the observed network structure of the posts [175], and introduced an approach for representing the structural dependencies of debate dialogues using graphical models and joint relational embeddings [117]. In contrast, our work captures relational structures to understand the context of the discussion and classify stances based on pairs of posts/sentences.

6.3 Model

To illustrate our model, we use some highly simplified notation. Let $F = \{f_1, f_2, \dots, f_N\}$ and $T = \{T_1, T_2, \dots, T_M\}$ respectively denote the sets of forum participants and topics, where each f_c represents a forum participant and each topic T_i consists of a set of sentences $\{s_{i,1}, s_{i,2}, \dots, s_{i,m_i}\}, \forall c \in \{1, \dots, N\}$ and $i \in \{1, \dots, M\}$. Each sentence s_j of topic T_i is mapped to its author f_c , and $\alpha_c^{T_i}$ represents the set of sentences belonging to T_i written by f_c . A forum participant f_c participates in a topic T_i if and only if $s_j \in \alpha_c^{T_i}$. The sentence s_j is the stance expressed by the forum participant f_c at time t and the sentence s_l is the stance expressed by the same person at time $t+n$ ($n > 0$)

6.3. Model

Table 6.1 – An example to demonstrate the functioning of our approach. We suppose a discussion forum of 10 sentences that involves three forum participants, $F=\{f_1, f_2, f_3\}$, who debate on a given topic. The sentence s_1 is the root of the discussion; it does not depend on a premise. It should be noted that the class labels of the couples are extracted separately using RoBERTa.

(a) Sentences by f_c		(b) Stance combinations
F	Sentence	Sentence pairs (premise, hypothesis)
f_1	s_1	–
f_2	s_2	(s_1, s_2)
f_3	s_3	$(s_1, s_3), (s_2, s_3)$
f_2	s_4	$(s_1, s_4), (s_2, s_4), (s_3, s_4)$
f_1	s_5	$(s_1, s_5), (s_4, s_5), (s_3, s_5)$
f_3	s_6	$(s_5, s_6), (s_4, s_6), (s_3, s_6)$
f_1	s_7	$(s_5, s_7), (s_4, s_7), (s_6, s_7)$
f_3	s_8	$(s_7, s_8), (s_4, s_8), (s_6, s_8)$
f_2	s_9	$(s_7, s_9), (s_4, s_9)$
f_1	s_{10}	(s_7, s_{10})

in the course of the debate. The sentence s_k denotes the stance expressed by another forum participant before s_l is voiced ($j < k < l$).

Respecting the timestamps of each sentence, we follow the flow of the discussion and learn the language inference between sentences to determine the stance class. To this end, we make combinations of sentence pairs between s_j and s_l . The rationale behind this is to identify the relationship between the meanings of a sentence pair by verifying whether s_k agrees with s_j , contradicts it or is simply neutral vis-a-vis the topic (T_i) that is being discussed. Specifically, we consider s_j as the premise and all possible s_k as the hypothesis for deciding the stance class. Let L denote a set of stance classes in the discussion that characterizes the position of the sentence s_k towards the sentence s_j , that is, the couple (s_j, s_k) : the premise-hypothesis relationship. The class of the couple (s_j, s_k) corresponds to z , $\forall z \in L$ and $\forall s_j, s_k \in T_i$, where $L=\{\text{Neutral}, \text{Agreement}, \text{Disagreement}\}$ and $k \in]j, l[$. A sentence without a predecessor could be considered as the root of the discussion and might not directly depend on any premise. We therefore lack couples for such cases.

To perform the sentence-pair classification task, we utilize RoBERTa [122], which is already fine-tuned on the MultiNLI (Multi-genre Natural Language Inference) corpus. The MultiNLI is a crowdsourced collection of 433K sentence pairs annotated with textual entailment information [199]. To classify two sentences, RoBERTa gen-

CHAPTER 6. A PRE-TRAINING APPROACH FOR STANCE CLASSIFICATION IN ONLINE FORUMS

Table 6.2 – An example to illustrate feature extraction. Suppose that s_1 and s_2 are respectively annotated beforehand as Agreement and Disagreement in Table 6.1(a) and RoBERTa yields a Neutral for the couple (s_1, s_2) . Clearly, s_1 represents the premise and s_2 the hypothesis. Therefore, we take Agreement as the premise label and Disagreement as the hypothesis label.

Sentence pair	Premise	Hypothesis
(s_1, s_2)	s_1	s_2
(s_1, s_3)	s_1	s_3
(s_2, s_3)	s_2	s_3
(s_1, s_4)	s_1	s_4
(s_2, s_4)	s_2	s_4
(s_3, s_4)	s_3	s_4
(s_1, s_5)	s_1	s_5
(s_4, s_5)	s_4	s_5
(s_3, s_5)	s_3	s_5
(s_5, s_6)	s_5	s_6
(s_4, s_6)	s_4	s_6
(s_3, s_6)	s_3	s_6
(s_5, s_7)	s_5	s_7
(s_4, s_7)	s_4	s_7
(s_6, s_7)	s_6	s_7
(s_7, s_8)	s_7	s_8
(s_4, s_8)	s_4	s_8
(s_6, s_8)	s_6	s_8
(s_7, s_9)	s_7	s_9
(s_4, s_9)	s_4	s_9
(s_7, s_{10})	s_7	s_{10}

erates fixed-size sentence embeddings where the feature representations of sentences are obtained from the trained encoders, and then passes them to a softmax classifier to derive the final label: i.e., *contradiction*, *neutral* or *entailment*. We obtain (i) entailment when the hypothesis has a similar meaning to the premise, (ii) contradiction when the hypothesis has a contradictory meaning, and (iii) neutral when the hypothesis has mostly the same lexical items as the premise but bears a different meaning. In this paper, we chose to use Agreement and Disagreement to simplify the terms entailment and contradiction, respectively.

Table 6.1 illustrates the functioning of our approach. We assume that each sentence in Table 6.1(a) solely addresses a single topic (T_1) which is argued by three forum participants. Table 6.1(b) shows how we generate possible combinations of sentence pairs based on the logic described above. To graphically represent the flow of the discussion, we take the relational structure depicted by Table 6.1(b) to construct

6.4. Experiments

topic-based graphs. More formally, $G = (V, A, T_i)$ denotes a directed multigraph for the topic T_i , where V is the set of vertices corresponding to forum participants and A is the set of arcs indicating stance labels. Recall that stance labels are results of sentence-pair classification yielded by RoBERTa, and a directed multigraph may have several arcs with the same origin and destination vertices. We explore G to measure the divergence of opinion throughout the discussion.

To classify stances, we design additional features from the dataset annotations, namely the premise and hypothesis labels. We assume that sentences in the dataset are annotated beforehand. For each couple (s_j, s_k) , we derive the sentence-pair label z using RoBERTa ($z \in L$). Additionally, we collect the true labels of these sentences as annotated, and then follow the position of each sentence in the couple to properly assign premise and hypothesis labels to them (see Table 6.2).

6.4 Experiments

To empirically evaluate our method, we conducted extensive experiments with two publicly available online forum datasets: Annotated Coarse Discourse and Internet Argument Corpus v2. We will now describe these datasets, introduce the techniques used as a baseline for comparison, and present the evaluation metric and details of the training process for our method.

Data Description. The Internet Argument Corpus v2 (IAC2) dataset is a collection of corpora of political debate topics on online forums [1]. Initially, it should be noted that IAC2 is composed of three different datasets: 4forums, which comprises over 3.5K participants and 414K posts (with an average of 340 users per topic and 19 posts per user); ConvinceMe (65K posts) and CreateDebate (3K posts). Of these, we opted to use 4forums because of its size and the number of users it contains, as well as for its topic annotations and response characterization. Notably, 4forums has crowdsourced annotations with a high inter-annotator agreement for stances of users in each topic and disagreement between users who reply to one another, and it spans many topics. In our experiments, we limited ourselves to five topics, Evolution, Gay Marriage, Abortion, Gun Control, and Death Penalty. On 4forums,

CHAPTER 6. A PRE-TRAINING APPROACH FOR STANCE CLASSIFICATION IN ONLINE FORUMS

agreement/disagreement scores are given on an 11-point scale $[-5,5]$. Scores ≤ 0 indicate higher inter-annotator confidence for disagreement, whereas scores ≥ 1 denote agreement. Sridhar et al. [175] removed all posts for which annotations belong to the interval $[0,1]$ due to uncertainty about the agreement. In contrast, we opted to keep these posts, since RoBERTa is fine-tuned to detect cases where the stance of the hypothesis sentence is neutral.

The Annotated Coarse Discourse (ACD) dataset is a large corpus of discourse annotations and relations collected from Reddit by [215]. Its goal is to allow a better understanding of online discussions at scale. It contains over 61K participants and 9,000 threads comprising over 101,000 comments, manually annotated. Basically, the discourse-act annotation scheme was developed to highlight comments that include agreement, appreciation, disagreement and negative reactions. In contrast to IAC2, we assume that ACD solely covers one topic, given that it does not include topic annotations.

Model Evaluation. To validate the performance of the proposed method, we compared it with the following baseline methods: PSL [175], UWB [102], MITRE [212], SRL [117] and BERT [43]. It should be recalled that we plainly discussed some limitations of the first four methods in Section 6.2.

- PSL [175] uses probabilistic soft logic to capture relational information in the network of authors and posts. The intuition of PSL is that the class Agreement or Disagreement between users correlates to their stance towards a topic.
- UWB [102] is based on a maximum entropy classifier with mainly surface-level, sentiment and domain-specific features.
- MITRE [212] maximizes the value of limited training data by transferring features from other systems trained on large, unlabeled datasets.
- SRL [117] uses the structural dependencies of the discussion and measures the similarity between embedding representations of the post and a given stance label.
- BERT [43] is a bidirectional Transformer-based pre-trained contextual representation trained using masked language modeling objective and next sentence

6.4. Experiments

prediction tasks. It exploits a multi-layer bidirectional Transformer encoder, where each layer contains multiple attention heads. More specifically, we utilize a BERT-large model fine-tuned on the MultiNLI [43, 92].

Feature sets. To build our stance classifiers, we used four different features. The feature `sentence-pair_label` is our response variable and refers to the stance label yielded by RoBERTa. This portrays the stance label of a sentence pair, i.e., the stance of a given hypothesis sentence towards a premise sentence. The features `premise_label` and `hypothesis_label` stem from human-annotated labels in the experimental datasets: we consider the manually annotated stance label for each sentence as ground truth. Finally, the feature `topic` denotes the topics discussed by the forum participants.

Model performance. To evaluate model performance, we conducted stratified ten-fold cross-validation to split our training and testing sets. We trained three distinct classifiers: logistic regression (LR), support vector machine (SVM) and random forest (RF). For SVM, we set the value γ of the radial basis function kernel to 0.5 and for RF, we built a model with 100 trees. We replicated the same logic for BERT to generate features, then trained an SVM classifier, BERT+SVM. We computed the F-1 score (harmonic mean of precision and recall) to measure the accuracy of our classifiers and to quantitatively compare them with the baseline techniques.

Divergence metrics. To quantify divergence of opinion within topic debates, we used four measures of probability divergence: Kullback–Leibler (KL), Jensen-Shannon (JS), Hellinger distance (HD) and Bhattacharyya distance (BD). These divergence metrics measure the discrepancy/similarity between two probability distributions [22, 99, 103, 165, 167].

Let us look at two discrete probability distributions $P=\{p_i\}_{i\in[n]}$ and $Q=\{q_i\}_{i\in[n]}$ supported on $[n]$. KL is a directed divergence that measures the discrepancy between the two, with the meaning being dependent on which direction was computed (see Eq. 6.1). Eq. 6.1 determines how the Q distribution is different from the P distribution. KL is a non-negative, asymmetric distance (i.e., $\text{KL}(P\|Q) \neq \text{KL}(Q\|P)$); it is zero if the two distributions are identical and can potentially equal infinity [167].

CHAPTER 6. A PRE-TRAINING APPROACH FOR STANCE CLASSIFICATION IN ONLINE FORUMS

JS is a symmetrized, smoothed version of KL which measures the total KL divergence from the average mixture distribution, $M = \frac{(P+Q)}{2}$ (see Eq. 6.2). Some salient features of JS are that it is always defined, bounded and symmetric, and only vanishes when $P=Q$ [22]. HD is a probabilistic analog of the Euclidean distance and satisfies the triangle inequality. The $\sqrt{2}$ in Eq. 6.3 is to ensure that $\text{HD}(P, Q) \leq 1$ for all probability distributions. One advantage of HD is that it serves to provide the lower bounds for Bayes risk in non-regular situations [165]. BD is defined as the negative logarithm of the Bhattacharyya coefficient [99]. Clearly, BD does not satisfy the triangle inequality, $0 \leq \text{BD}(P, Q) \leq +\infty$ (see Eq. 6.4). The Bhattacharyya measure has a simple geometric interpretation as the cosine of the angle between two position vectors in n -dimensional space $(\sqrt{p_1}, \dots, \sqrt{p_n})^\top$ and $(\sqrt{q_1}, \dots, \sqrt{q_n})^\top$, where $\cos(\theta) = \sum_{i \in [n]} \sqrt{p_i \times q_i}$. Consequently, if the P and Q distributions are identical, $\cos(\theta) = 1$, corresponding to $\theta = 0$.

$$\text{KL}(P\|Q) = \sum_{i \in [n]} p_i \times \log\left(\frac{p_i}{q_i}\right), \quad (6.1)$$

$$\text{JS}(P\|Q) = \frac{1}{2}\text{KL}(P\|M) + \frac{1}{2}\text{KL}(Q\|M), \quad (6.2)$$

$$\text{HD}(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i \in [n]} (\sqrt{p_i} - \sqrt{q_i})^2} \quad \text{and} \quad (6.3)$$

$$\text{BD}(P, Q) = -\ln\left(\sum_{i \in [n]} \sqrt{p_i \times q_i}\right). \quad (6.4)$$

We assume that one is more likely to encounter divergences of opinion in sentence pairs for which the class label is `Disagreement` or `neutral`. We therefore take each of these sentence pairs, apply the word2vec skip-gram model to embed the words of each sentence of the pair as vectors in a low-dimensional space [129], and finally encode them as probability densities [20]. Note that the densities represent the distributions over the possible significations of a word. We calculate the divergence metrics between the distributions of sentence pairs. To estimate divergence of opinion on the whole topic, we calculate the arithmetic mean of the results obtained from all sentence pairs of interest.

6.5. RESULTS AND DISCUSSION

Table 6.3 – Stance classification results for the proposed method and baselines on the two experimental datasets. The F-1 score metric is used to gauge model performance. Bold font indicates the best results for each class label. PSL, UWB, MITRE and SRL merely classified two classes, i.e., Agreement and Disagreement.

	Label	Ours+LR	Ours+SVM	Ours+RF	BERT+SVM	PSL	UWB	MITRE	SRL
IAC2	Agreement	0.803	0.796	0.788	0.781	0.572	0.687	0.756	0.671
	Disagreement	0.807	0.812	0.746	0.737	0.56	0.675	0.753	0.664
	Neutral	0.799	0.805	0.814	0.795	–	–	–	–
ACD	Agreement	0.787	0.772	0.768	0.762	0.495	0.648	0.74	0.626
	Disagreement	0.751	0.778	0.785	0.746	0.496	0.654	0.744	0.631
	Neutral	0.776	0.767	0.780	0.764	–	–	–	–

6.5 Results and discussion

Stance classification. Table 6.3 presents the performance results for three-class classification in different experimental settings. It should be noted that the baseline techniques only classify two classes (Agreement and Disagreement), except for BERT+SVM which classifies the three classes (Agreement, Disagreement and neutral). Our classifiers and BERT+SVM are able to accurately distinguish between the three classes and achieve better performance than the other baseline techniques.

We observe that the F-1 scores are significantly higher than 0.5 for stance classification, although PSL achieved poor performance on ACD. However, this poor performance can be partly explained by PSL’s inability to capture the context of the discussion. Our method yields statistically significant improvements over PSL, surpassing it by 0.292 for Agreement and 0.289 for Disagreement on ACD. We empirically show that our method outperforms the baselines by a considerable margin for the 2-classes and yields good classification performance on the label neutral. Further investigation found that the arithmetic mean of our method’s F-1 scores on the 2-classes also surpassed the baselines (with 0.796 for Agreement and 0.788 for Disagreement on IAC2, and 0.776 for Agreement and 0.771 for Disagreement on ACD). We note that UWB and SRL achieved roughly similar performances on the two experimental datasets.

Compared with MITRE, BERT+SVM attained better performance on both datasets,

CHAPTER 6. A PRE-TRAINING APPROACH FOR STANCE CLASSIFICATION IN ONLINE FORUMS

Table 6.4 – Quantifying divergence of opinion by topic.

	Topic	JS	KL	HD	BD
IAC2	Evolution	0.114	0.089	0.103	0.077
	Gay Marriage	0.206	0.193	0.198	0.185
	Abortion	0.337	0.323	0.305	0.294
	Gun Control	0.442	0.437	0.421	0.405
	Death Penalty	0.253	0.268	0.244	0.231
ACD	Coarse	0.414	0.406	0.409	0.382

but still lagged behind our classifiers. Specifically, the average of our best performances yields F-1 scores higher than those for the strongest baseline, BERT+SVM, with improvements of about 0.039 and 0.027 over IAC2 and ACD, respectively. BERT+SVM achieved performances closer to those of our smallest classifier for the class `neutral`: 0.004 on IAC2 and 0.003 on ACD. It should be noted that, aside from BERT+SVM, MITRE is the baseline that comes closest to our best F-1 scores. We observe that our method improves upon MITRE by (0.047, 0.059) and (0.047, 0.041) for Agreement and Disagreement on (IAC2, ACD), ($p < 0.0005$ as per the McNemar test [127] on IAC2 and $p < 0.0003$ on ACD). MITRE performs similarly on the 2-classes on both datasets, and once outperformed our smallest classifier and BERT+SVM on the class Disagreement, i.e., (BERT+SVM < Ours+RF < MITRE < Ours+LR < Ours+SVM). Note that MITRE is a transfer-learning method trained on large unlabeled datasets to generate features using word embeddings, and then learn sentence representations from these features to classify stances. MITRE retains the knowledge acquired in solving one case and subsequently applies it to a different but related case. This explains its good performance.

We have shown clear benefits and strong evidence that capturing the context and relational structures of debates can provide better performance on the task of stance classification. We achieved our best F-1 scores on IAC2 with LR and on ACD with RF. The highest F-1 score, 0.814, was achieved for the class `neutral` on IAC2, and the smallest F-1 score obtained by the proposed method is much greater than the F-1 scores of all baseline techniques except for MITRE; that is, (MITRE > Ours+RF > BERT+SVM > UWB > SRL > PSL).

Divergence within debate topics. Table 6.4 shows the results of the divergence

6.5. RESULTS AND DISCUSSION

metrics we utilized to measure the divergence of opinion on each debate topic addressed in the forum. The metrics used yielded approximately similar results, even though JS achieved the best performance on the majority of topics tackled in the experimental datasets. We note that the values for divergence of opinion yielded by the experimental metrics do not exceed 0.50. The significance of the overall divergence value may be somewhat difficult to interpret, whereas divergence of opinion between two different viewpoints can be understood and explained. However, analyzing the motives and sentiments behind divergences of opinion which fueled heated discussions normally requires further inquiry.

We observe that `Gun Control` is a topic that sparked a relatively large divergence of opinion between proponents and opponents of the right to keep and bear arms, and yielded a higher divergence value than other topics on IAC2 (with 0.442 over JS and 0.426 as the arithmetic mean of the divergence values of all metrics). We notice that KL and HD performed similarly on `Gay Marriage` on IAC2. We found that `Evolution` is the topic with the smallest divergence value, followed by `Gay Marriage`. Moreover, the topic `Abortion` generated a greater divergence of opinion than `Evolution` and `Gay Marriage` combined ($0.337 > 0.32$ with JS, $0.323 > 0.282$ with KL, $0.305 > 0.301$ with HD, $0.294 > 0.262$ with BD).

We note that KL achieved better performance than JS on `Death Penalty` ($0.268 > 0.253$). (Proponents of this topic argue that capital punishment is beneficial even if it has no deterrent effect, while opponents alternatively suggest life imprisonment.) We obtained an arithmetic mean of 0.196 on the whole discussion on `Gay Marriage`. Somehow, this value is greatly inferior to that for cases where there is no divergence (a difference of $1 - 0.196$), i.e., 0.804, suggesting that this topic may have triggered relatively few emotional debates.¹¹ Finally, we observe that JS and BD achieved the highest and lowest divergence values, respectively, on ACD.

11. Based on the results yielded in Table 6.5, we find that the correlation of negative emotion-related words over `Gay Marriage` is not statistically significant. This could be considered as strong evidence to argue that this topic may have sparked few emotional statements.

CHAPTER 6. A PRE-TRAINING APPROACH FOR STANCE CLASSIFICATION IN ONLINE FORUMS

Table 6.5 – Prediction performance (Pearson’s r) based on 10-fold cross-validation using LIWC features (positive and negative emotions) extracted from different topics addressed on IAC2 and ACD datasets. All features are significant at $p < 0.005$, except for the negative emotion on Gay Marriage and the positive emotion on Coarse, for which p is not statistically significant.

LIWC	Evolution	Gay Marriage	Abortion	Gun Control	Death Penalty	Coarse
Positive emotion	0.301	0.225	0.26	0.425	0.218	0.151
Negative emotion	0.287	0.163	0.271	0.409	0.43	0.467

6.6 Psychological processes

To measure emotional state manifested [57] by forum participants who addressed the topics that we studied above, we utilize Linguistic Inquiry and Word Count (LIWC) [144], a dictionary which is widely employed in computational linguistics as a source of features for psychological and psycholinguistic analysis. LIWC comprises words that have very clear, pre-labeled meanings. The dictionary includes words in various categories, notably linguistic dimensions, psychological processes and personal concerns. Each category is found to be correlated with several psychological traits and outcomes [65, 66]. Within the psychological processes category, we find the emotion sub-dictionaries, that is, positive and negative emotions. We focus on the psychological processes category in order to explore the linguistic usage in user viewpoints. It should be noted that the positive and negative emotions are not two ends of a scale, since a point of view can include the two. We leverage each opinion (each sentence, see Table 6.2) and measure the proportion of word tokens that fall into negative and positive emotions.

To predict the emotion associated with opinions by topic, we treat each topic separately and stratify each topic’s data by 10-fold cross-validation to split our training and testing sets. We utilize linear regression with three different regularization methods: LASSO, ridge and elastic net. The elastic net yielded marginally higher performance over the two other techniques. The performance was measured using the Pearson correlation (r) [143]. Table 6.5 indicates the correlations between words containing positive and negative emotions over different topics addressed in the two experimental datasets. It can be seen that Coarse yielded the strongest correlation

6.7. Conclusion

with negative emotion, 0.467 ($p < 0.001$) and `Gun Control` produced the highest correlation with positive emotion, 0.425 ($p < 0.001$). More importantly, we find that both positive and negative emotion features for all topics are significant at $p < 0.005$, apart from the p for the negative emotions on `Gay Marriage` ($p = 0.756$) and positive emotions on `Coarse` ($p = 0.725$) and `Death Penalty` ($p = 0.618$), respectively, which are not statistically significant. We observe that positive emotions appear to have a stronger effect on `Gay Marriage` ($r = 0.225$, $p < 0.001$) and `Evolution` ($r = 0.301$, $p < 0.001$).

6.7 Conclusion

We present a RoBERTa-based method to classify stances by capturing the context and relational structures of the debate. Our method shows statistically significant improvements over existing methods in terms of F-1 score performance on this task and provides good results for the class `Neutral`, which was not considered in prior work. We report that `Neutral` yields performance surpassing 0.75 on the two experimental datasets. Furthermore, we examine the degree of disagreement and neutrality to measure the divergence of opinion on topics addressed in the debate. We note that none of the metrics utilized yields values surpassing 0.5. We limit ourselves to reporting the observed divergence values rather than explaining the motives and sentiments that fueled the debate so that we have divergence of opinion among individuals; this aspect normally requires further analysis. We measure the emotional state manifested in topics addressed in different debates. To this end, we resort to the LIWC dictionary, especially the psychological processes category, to calculate the proportion of opinion-related words that fall into positive and negative emotions. We find that the majority of features extracted from all topics addressed are statistically significant. Additionally, we indicate the topics addressed that include the highest and lowest correlations of positive and negative emotions.

This study provides a framework for further research about stance classification in different settings in online discussion forums. Specifically, we aim to exploit pre-trained language models to classify stances based on hypotheses related to multiple independent premise sentences [108] and thereafter detect some logical fallacies in

CHAPTER 6. A PRE-TRAINING APPROACH FOR STANCE CLASSIFICATION IN ONLINE FORUMS

debates, including strawman, red herring, *tu quoque*, hasty generalization and slippery slope arguments. Furthermore, we would like to study the effects of emotional reactions on divergent opinions, investigate at the user- and debate-levels in order to discern the motives behind divergent opinions, and predict whether the intensity of emotional reaction in divergent opinions is likely to grow as the debate moves forward.

Chapter 7

On Predicting Behavioral Deterioration in Online Discussion Forums

Jean Marie Tshimula,¹ Belkacem Chikhaoui,^{1,2} Shengrui Wang¹

¹Département d'informatique, Université de Sherbrooke, QC J1K 2R1, Canada

²LICEF Research Center, Université TÉLUQ, QC H2S 3L5, Canada
{kabj2801, shengrui.wang}@usherbrooke.ca, belkacem.chikhaoui@teluq.ca

Keywords: Misbehavior, Behavioral sequences, Deterioration.

Abstract

Early detection of behavioral deterioration can be of great importance in preventing individuals' misbehavior from escalating in severity. This paper addresses the problem of behavioral deterioration in the context of online discussion forums. We propose a novel method that builds behavioral sequences from temporal information to gain a better understanding of behaviors exhibited by forum members, and then explores n -gram features to predict behavioral deterioration from consecutive combinations of sequential patterns corresponding to misbehavior. We conduct extensive

CHAPTER 7. ON PREDICTING BEHAVIORAL DETERIORATION IN ONLINE DISCUSSION FORUMS

experiments using real-world datasets and demonstrate the ability of our method to predict behavioral deterioration with a high degree of accuracy, as evaluated by F-1 scores. Our quantitative analysis of the model’s performance yields F-1 scores of over 0.7. Specifically, we find that the best-performing model is linear SVM, with an average F-1 score of 0.74. Some future research avenues are proposed.

7.1 Introduction

The advent of online forums has revolutionized the speed of world connectivity, real-time information sharing, information discovery, real-time news, and instant communication, and creates new possibilities for investigating user behaviors through their digital footprints. Online forums aim to nurture social behavior, a sense of community and affinity relationships among individuals [185, 186]. Increasingly, however, they are having the opposite effect, due to a rising tide of deviations and deliberate provocations. While some people show common sense, tolerance and respect for the views of other forum members, others manifest intransigent attitudes and engage in misbehavior that harms the community and adversely affects the equanimity of forum members. The safety, usability, and reliability of online discussion forums may thus be compromised due to the prevalence of abuse and misbehavior expressed in various ways, such as videos, pictures, taunting emoticons and comments, to just name a few. In this paper, we limit our investigation to textual data and assemble different classes of temporal behavior displayed by individuals into more interpretable sequences.

Misbehavior may refer to disruptive acts characterized by covert or overt hostility and intentional aggression towards others [31, 70, 84, 87, 104]. There is substantial evidence that the display of aggressive emotions is a valid predictor of risk factors for violence [8]. People who engage in misbehavior may severely transgress against social norms and social expectations for a particular environment, including full participation, right to safety and privacy, right to freedom of opinion and expression, decency, etc. Covert hostility can be expressed in one-to-one or one-to-many communication, whereas overt hostility can be voiced in online forums [104, 163]. It

7.1. Introduction

should be noted that misbehavior includes but is not limited to abusive and offensive language, threats, hate speech, cyberbullying, and race and gender discrimination [18, 39]. Waseem et al. [197] studied how these behaviors are related and proposed a typology that captures the similarities and differences among them. This provides a ground truth for predicting future behavior with sufficient certainty.

Recent research has reported descriptive statistics on the number of victims of misbehavior. Kumar et al. [105] found that 40% of Internet users had experienced cyberbullying. Blumenfeld and Cooper [18] reported that 54% of LGBT youth had been cyberbullied. Li [118] found that nearly 54% of students were victims of traditional bullying and over a quarter of them had been cyberbullied. Additionally, their study found that roughly 60% of cybervictims were female and 39% were male. Waldman and Verga [194] put forward that 90% of terrorist activities on the Internet are conducted within online social networks. Some instances of misbehavior may initially have small statistical effects, but their persistent accumulation may subsequently have major and devastating consequences. Persistent misbehavior is a proven risk factor for a number of serious problems. For example, some victims of cyberbullying are more likely to self-harm, engage in suicidal behavior [94], and experience some unpleasant aftermaths, including psychological and anxiety disorders [40, 87, 120]; others even commit suicide [79].

Evidence from the research discussed above shows a tremendous need for efficient approaches capable of preemptively detecting misbehavior as early as possible. In the absence of such approaches, misbehavior can escalate to violent behavior when the perpetrators constantly harm other forum members and do not get sanctioned for their misdeeds. Violent behavior may thus be considered as the endpoint on a continuum of behavioral deterioration [52]. Behavioral deterioration may occur suddenly or slowly, depending upon the pace at which perpetrators cause harm. More specifically, we define deterioration as the accumulation of misbehavior.

The detection of misbehavior can be quite challenging and complex, for several practical reasons. Different people may have different ways of expressing the same misbehavior: for instance, masked pejorative terms, more subtle biases, coded messages and/or figures of speech (such as metaphor) may be used to misrepresent disparate impact [39, 161]. Recently, Mozafari et al. [133] introduced a BERT-based

CHAPTER 7. ON PREDICTING BEHAVIORAL DETERIORATION IN ONLINE DISCUSSION FORUMS

misbehavior classifier. This system suggests new fine-tuning strategies to investigate the effect of different layers of BERT and shows the ability to take contextual information into account, capture various ways in which misbehavior is expressed, and classify misbehavior classes more efficiently. In this paper, we resort to this model for building behavioral sequences from temporal behaviors exhibited by forum members in order to predict behavioral deterioration. To the best of our knowledge, our paper is the first to address the problem of behavioral deterioration in the context of online discussion forums.

Specifically, the key contributions of this paper can be summarized as follows:

- We first introduce a formal definition of the problem of behavioral deterioration.
- We then propose a method that constructs behavioral sequences from consecutive combinations of misbehavior classes and explores n -gram features to gain a better understanding of behavior exhibited by forum members and predict behavioral deterioration over time.
- We conduct extensive experiments using two publicly available datasets to validate the behavioral deterioration prediction. Our method is conceptually simple and highly interpretable.

The remainder of this paper is organized as follows. In Section 7.2, we discuss some related work and the rationale for detecting signals relevant to deterioration. Section 7.3 describes the proposed method and the feature set extracted to train predictive models with alternative combination of feature sets. We present experiments in Section 7.4. Section 7.5 is devoted to the discussion of our outcomes and the limitations of the study. Finally, we present our conclusions and propose future research directions in Section 7.6.

7.2 Related work

Topic-based user behavior. Gong and Wang [59] introduced a holistic user behavior modeling approach to understand user intentions, relying on both sentiment and social network analysis to collect behavior patterns for each user. They developed a probabilistic generative model incorporating two learning tasks—opinionated content

7.2. Related work

modeling and social network structure modeling—to recognize user preferences and their relatedness, respectively. In the first task, logistic regression is utilized to map sentiment polarity from textual content generated by a statistical language model based on a v -dimensional multinomial distribution over the vocabulary (v denotes the vocabulary size). In the second, a stochastic block model is employed to capture the relatedness among users. Wang et al. [196] explored the first task and proposed an unsupervised neural network-based model to learn linguistic descriptors for the user’s behavior over time. The method discovers linguistic dissimilarities that correlate with user activity levels and community clustering. While correlation does not imply causation, Aumayr and Hayes [11] sought to depict the correlation between clustered behaviors and three predefined topic properties (accessibility, sociability, and controversy). Their rationale was to present the effects that certain sorts of topics may have on user behavior, although the cluster categories were manually labeled to make the dendrogram more explicit. Furthermore, user behaviors were drawn from topics that they participated in rather than from opinions they expressed in the forum. We assume that this may result in failure to capture some signals that could be relevant to deterioration.

Hassan et al. [73] introduced a method for detecting the attitude of users towards others. Their approach involves training a supervised Markov model of the lexical item, part-of-speech tags, and dependency patterns to build a model capable of identifying sentences with and without attitude. Along similar lines, Zhai et al. [214] proposed an unsupervised approach based on the evaluation of opinion sentences to remove those which contain emotional statements, personal attacks and opinions that do not express positive views about the discussion topics. Zhang et al. [216] detected early signs of conversational failures, such as harassment and personal attacks. More recently, Cliche [38] introduced a deep-learning-based classifier to tackle sentiment analysis issues. Their classifier leverages a large quantity of unlabeled information, using 100 million unlabeled tweets to pre-train word embeddings via distant supervision before applying convolutional neural networks and an attention-based biLSTM approach for classifying noisy positive and noisy negative tweets. We note some limitations of the aforementioned research, including the inability to verify whether individuals keep expressing opinions with or without attitude over time. In contrast

CHAPTER 7. ON PREDICTING BEHAVIORAL DETERIORATION IN ONLINE DISCUSSION FORUMS

to these studies, we examine the temporal behaviors exhibited by forum members and assemble them in behavioral sequences to predict whether their behavior is affable or tends to deteriorate.

Zhao et al. [217] proposed a behavioral factorization (BF) method to model behaviors of each user based on topic interests derived from publishing signals such as posts, shares, likes, etc. BF learns a latent embedding model by factoring matrices split into behaviors (behavior-non-specific user-topic, single behavior-specific user-topic, and combined behavior-specific user-topic matrices) and then builds user topic profiles for various behavior types using the latent embedding space. The limitation of this work is that it draws solely on discussion topics addressed by forum members and does not regard different types of behavior they displayed in their posts.

Malicious and aggressive behavior. Cheng et al. [31] detected users engaged in antisocial behavior that negatively impinges on other users and causes harm to the community, and predicted whether some users would be banned from the community based on their overall activities. Specifically, they compared the activities of users who have been banned in the past with those who have never been banned. To this end, the model deals with user posts, including data from features that allow users to upvote, downvote, report a post, etc. One limitation of this work is that the model relies more heavily on other features than on user posts to identify whether reported posts contain unpleasant statements. A post may be reported for the use of offensive language although the content of the post does not justify the accusations. The study does not address such cases. We believe that the model’s failure to deal with post content is a shortcoming, as relying only on abuse-report-based features may be misleading to some extent.

Razavi et al. [155] reported work on multi-level classifiers enhanced by an Insulting or Abusive Language Dictionary (IALD) they developed to detect offensive language in text messages. Two rule-based auxiliary tools are proposed. One is the rearmost level of the classifiers and the other is utilized for constructing patterns out of the IALD. Several solutions to the problems they address have been put forward in the literature, in particular for detecting cyberbullying, hate speech and offensive language in online communities [41, 75, 106, 124, 154, 159, 197, 203, 216]. In contrast

7.3. Model

to these studies, Mozafari et al. [133] proposed a BERT-based misbehavior classifier which outperforms several best-performing misbehavior classification techniques and understands and captures various ways in which misbehavior is expressed. We use this classifier [133] to construct behavioral sequences from temporal behaviors exhibited by individuals in order to predict behavioral deterioration.

7.3 Model

To develop our model, we utilize some simple notation. Let $S = \{s_1, s_2, \dots, s_K\}$ denote a sequence of K sentences in a forum, $F = \{f_1, f_2, \dots, f_H\}$ be a set of forum members, and $\alpha_S^{f_i}$ represent the set of sentences written by f_i in S , where $i \in \{1, \dots, H\}$. A forum member participates in the discussion if there is an l such that $1 \leq l \leq K$ and $s_l \in \alpha_S^{f_i}$. We assume that each such s_l is annotated beforehand in order to capture different types of behavior exhibited by f_i and facilitate behavior classification. Let $\beta_B^{f_i} = \{B_1, B_2, \dots, B_T\}$ be the set of behavior sequences exhibited by each forum member f_i , where $B_t = \{b_1^t, b_2^t, \dots, b_m^t\}$ and $t \in \{1, \dots, T\}$. Specifically, the sequence B_t represents the concatenation of all behavior label classes b_j^t exhibited by f_i in the period t , $\forall j \in \{1, \dots, m\}$ and $b_j^t \in \{N, M\}$. The classes N and M designate normal behavior and misbehavior, respectively. It should be noted that the b_j^t are derived using a classifier.

To perform behavior classification, we use the BERT-based misbehavior classifier introduced in [133]. Fundamentally, BERT is a recent Transformer-based pre-trained contextualized embedding model extendable to a classification model with an additional output layer [43, 133]. It has yielded state-of-the-art results on numerous benchmarks, including text classification and language inference, without substantial task-specific modifications. The rationale behind the BERT-based misbehavior classifier [133] is that it exploits new fine-tuning strategies to capture different levels of syntactic and semantic information, and this enables it to consider tiny details in texts and to perceive different ways in which misbehavior is expressed. The contributions of this method are briefly discussed in [133].

Suppose that f_1 exhibits the behavior sequence NMMMMNMMMMNM in the period t . The period is the interval of time elapsed between two different timestamps. We

CHAPTER 7. ON PREDICTING BEHAVIORAL DETERIORATION IN ONLINE DISCUSSION FORUMS

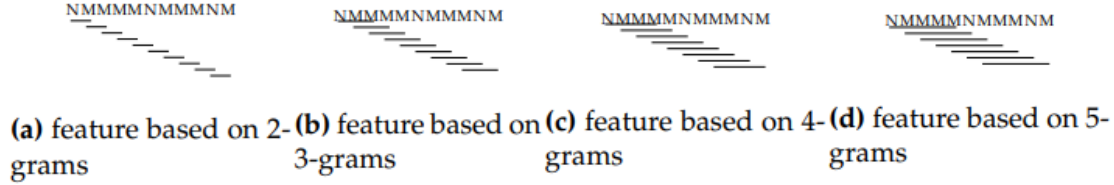


Figure 7.1 – Feature extraction process used to capture deterioration patterns within behavioral sequences (BS). Patterns are extracted based on all possible n -grams in BS from left-to-right. This allows us to better discern accumulations of behavior classes.

assume that deterioration cues can be observed from the accumulation of misbehavior classes.

To explore behavioral sequences, we design character n -gram features in order to capture signals that are potentially relevant to deterioration. The n -gram features with a pair of values (h_k, v_k) are extracted as input signals to be fed to a classifier. Specifically, h_k represents the n -gram feature k and v_k denotes the count of the feature within behavioral sequences. The n -grams can be generated by sliding a window of length n over the sequence B_t . Fig. 7.1 illustrates how n -gram features can be extracted from behavioral sequences. For instance, the features extracted from the behavioral sequence above can be presented as follows: 2-grams $\{(\text{NM}, 3), (\text{MN}, 2), (\text{MM}, 5)\}$, 3-grams $\{(\text{MMM}, 3), (\text{NMM}, 2), (\text{MMN}, 2), (\text{MNM}, 2)\}$, 4-grams $\{(\text{MMMMN}, 2), (\text{NMMMM}, 2), (\text{MMNM}, 2), (\text{MMMM}, 1), (\text{MNMM}, 1)\}$ and 5-grams $\{(\text{MMMMNM}, 2), (\text{NMMMM}, 1), (\text{MMMMN}, 1), (\text{MMNM}, 1), (\text{MNMM}, 1), (\text{NMMMM}, 1)\}$.

To classify behavioral deterioration, we design four different features using n -grams of order 2, 3, 4 and 5, respectively (Fig. 7.1). We use the constructed features to train linear support vector machines (SVM) and logistic regression (LR) classifiers. Basically, we label n -grams that support the accumulation of misbehavior classes as Deterioration and other n -grams as Non-deterioration. It should be noted that 4- and 5-grams which do not fully support the accumulation of misbehavior classes are treated differently. We consider them as full-fledged behavioral sequences and investigate the trend of their sub-2-grams by applying the same logic as in Fig. 7.1(a). The choice of sub-2-grams is arbitrary. The principal reason for exploring

7.4. Experimental setup

sub-2-grams is to better track the momentum of the accumulation of different behavior classes and discover deterioration patterns. We label these 4- and 5-grams as *Deterioration* based on whether the majority of the sub-2-grams they contain support the accumulation of misbehavior classes. For instance, MNMM comprises $\{(MN, 1), (NM, 1), (MM, 1)\}$; NMMM, $\{(NM, 1), (MM, 2)\}$; MMMMN, $\{(MM, 3), (MN, 1)\}$; NMMMM, $\{(NM, 1), (MM, 3)\}$; and NMMMN comprises $\{(NM, 1), (MM, 2), (MN, 1)\}$. We therefore label them as follows: $\{NMMM, MMMMN, NMMMM\}$ as *Deterioration* and $\{MNMM, NMMMN\}$ as *Non-deterioration*.

7.4 Experimental setup

To empirically evaluate our method, we conducted experiments using two publicly available online discussion datasets: HatebaseTwitter [41] and TRAC [104].

Datasets. HatebaseTwitter is a collection of 24,802 tweets and contains three labels: hate, offensive, and neither. TRAC contains 15,869 Facebook comments labeled as overtly aggressive, covertly aggressive, and non-aggressive. To classify the class labels of experimental datasets, we applied the BERT-based misbehavior classifier [133]. This method outperforms [41] and [197] and yields accuracies of 96.2% and 94.8% on HatebaseTwitter and TRAC, respectively (versus 90% for [41] on HatebaseTwitter, and 80% and 89% for [197] on TRAC and HatebaseTwitter). We therefore took the predicted classes produced by [133] to design behavioral sequences on a weekly basis: i.e., each sequence represents behaviors exhibited by an online forum member in the course of the week. The choice of the period over which to form the behavioral sequence is arbitrary and depends on how one wants to learn the deterioration distribution. To better explore sequence variation and follow deterioration cues, we chose to simplify the sequence by converting all misbehavior-related classes into "M" and the normal behavior class into "N". The major reason for using binary classes is to explore the behavioral sequences with a small number of object types in order to examine them thoroughly. We therefore utilized the designed features and the two classifiers for experimental settings, as mentioned in Section 7.3.

Model evaluation. To evaluate the performance of our model, we used 10-fold

CHAPTER 7. ON PREDICTING BEHAVIORAL DETERIORATION IN ONLINE DISCUSSION FORUMS

Table 7.1 – Results of behavioral deterioration prediction. Bold font indicates the best results for each class label.

	Class	HatebaseTwitter	TRAC
Ours+SVM	Deterioration	0.722	0.785
	Non-deterioration	0.718	0.749
Ours+LR	Deterioration	0.72	0.761
	Non-deterioration	0.719	0.758
LSTM	Deterioration	0.719	0.737
	Non-deterioration	0.716	0.733

cross-validation to split our training and testing sets. We computed F-1 scores to measure the accuracy of our classifiers and quantitatively compared them with the baseline. We used long short-term memory (LSTM) [80] as the baseline since it deals very well with long sequences and captures long-term dependencies. Note that we did not find an existing approach for detecting behavioral deterioration in the context of online forums.

7.5 Results and discussion

We demonstrate that quantifiable signals relevant to accumulations of misbehavior classes can be used for behavioral deterioration prediction. Table 7.1 presents the performance results of our method and the baseline. We observe that the F-1 scores for our classifiers and LSTM are significantly higher and show the ability to predict behavioral deterioration, with F-1 scores of over 0.7 for both classes.

All classifiers showed significantly better results for the class *Deterioration*. Note that our method achieved higher F-1 scores on both datasets. The results of LSTM on HatebaseTwitter are not far behind, while on TRAC the differences widen by a considerable margin for both classes, especially evident in the values 0.048 and 0.024 for the class *Deterioration* with Ours+SVM and Ours+LR, respectively. It should be noted that Ours+SVM was the best-performing classifier, yielding an average F-1 score of 0.74, and Ours+SVM and Ours+LR achieve approximately the same results on HatebaseTwitter. Additionally, we note that Ours+LR performs in more balanced ways and remark that the differences between its predicted class labels are smaller than those yielded by Ours+SVM: $(0.001 < 0.004)$ on HatebaseTwitter and $(0.003 <$

7.5. Results and discussion

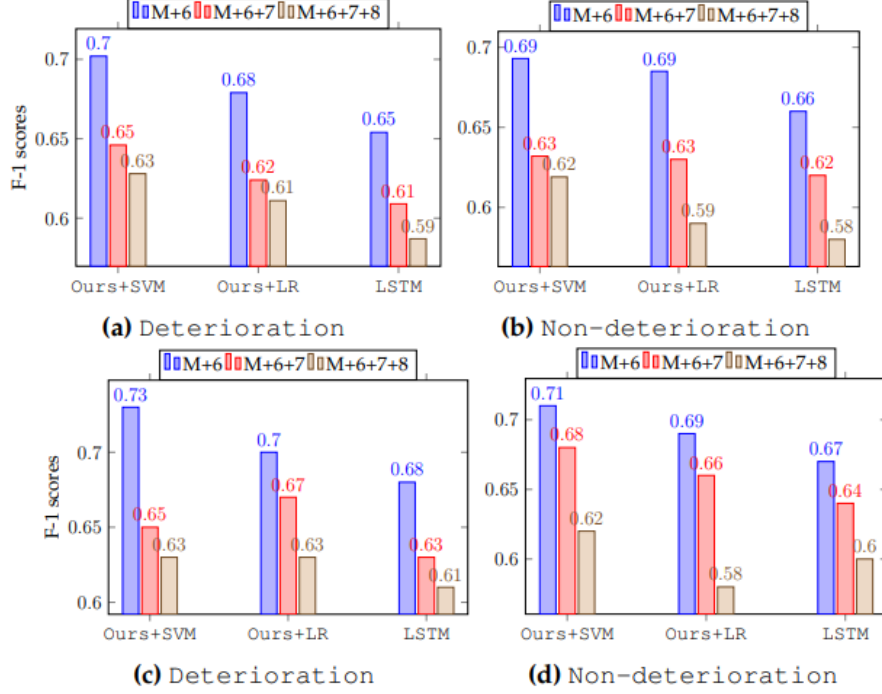


Figure 7.2 – Results of behavioral deterioration prediction by adding extra features to the main model. Specifically, M+6 means that we add 6-grams to the model, M+6+7 stands for 6- and 7-grams and M+6+7+8 means that we add 6-, 7- and 8-grams. We predict (a) and (b) on HatebaseTwitter and (c) and (d) on TRAC.

0.036) on TRAC.

To validate the performance of our models in generating deterioration estimates adequate to provide a good solution for particular individuals, some ground truth information is required. However, ground truth information to capture the prediction accuracy of behavioral deterioration is scarce and constitutes a challenging problem which has not been addressed in the context of online forums. It should be noted that the lack of ground truth information does not affect the generalizability of the findings and model performance, since the results stem directly from observed accumulations of behaviors exhibited by individuals in the discussion forum; and this makes intuitive sense. The number of features, as well as the number of elements in each n -gram,

CHAPTER 7. ON PREDICTING BEHAVIORAL DETERIORATION IN ONLINE DISCUSSION FORUMS

may be arbitrary and depend heavily on the average length of behavioral sequences. To explore the effect of the number of features on model performance, we extend the initially-built model by including in it some supplementary features to examine deterioration patterns within behavioral sequences. It should be recalled that the average length of the set of behavioral sequences that we constructed above is 9. Consequently, we add to the initially-built model: a feature extracted on 6-grams (M+6); two features extracted on 6- and 7-grams, respectively (M+6+7); and three features extracted on 6-, 7- and 8-grams, respectively (M+6+7+8). We treated differently 6-, 7- and 8-grams which do not fully support the accumulation of misbehavior classes by applying the same logic as for 4- and 5-grams, as described in Section 7.3.

Figure 7.2 presents the results of behavioral deterioration prediction with additional features. We report that the average performances yielded by our models exceed 0.6; that is, (0.691, 0.635, 0.619) for *Deterioration* and (0.689, 0.631, 0.605) for *Non-Deterioration* with M+6, M+6+7, and M+6+7+8, respectively, on HatebaseTwitter; and (0.715, 0.66, 0.63) for *Deterioration* and (0.7, 0.67, 0.605) for *Non-Deterioration* with M+6, M+6+7, and M+6+7+8, respectively, on TRAC. Our models achieved better results than LSTM on the two experimental datasets. We observe that the model performance decreases when the number of features increases. To examine deterioration patterns more closely, we suggest constructing a model based on $z-1$ features if the average length of overall behavioral sequences corresponds to z ($z > 2$). Following this logic, the model is supposed to utilize the feature sets varying from 2-grams to $(z-1)$ -grams. We assume that this renders it possible to extract longer accumulations of behavior classes to investigate deterioration patterns on various facets. Beyond monitoring accumulations of behavior classes to extract feature sets, we face challenges in defining threshold values (or early warning scores) to determine whether a set of behavioral sequences for individuals tends toward deterioration or not. Such scores could allow the establishment of different degrees of deterioration in order to facilitate more effective monitoring of the trajectory of behavioral deterioration. With thresholds fixed, we can identify deterioration at a sufficiently early stage to prevent significant further deterioration [56] and examine an individual’s mental state and personality traits [58, 148].

Our results provide strong evidence that we can predict behavioral deterioration

7.6. Conclusion

with an accuracy exceeding 0.6 (Table 7.1 and Figure 7.2), a resolution that is likely fine-grained enough for various experimental datasets. Significant signals relevant to deterioration remain to be uncovered and understood within behavioral sequences, including (i) examining correlations between the language use of individuals for which behavior sequences comprise accumulations of behavior classes that indicate signals relevant to deterioration; (ii) analyzing personality traits to understand whether deterioration occurs under the effects of the topics addressed in the discussion forum, or is related to mental health conditions or other reasons; (iii) understanding the impact of some personal concerns (such as work, money, religion, death, etc.) on behavioral deterioration and (iv) constructing a holistic model to explain the deterioration in relation to several factors at once [44, 110]. Developing these algorithms and evaluating them is a promising direction for future research.

7.6 Conclusion

We present a method that constructs behavioral sequences from forum members’ temporal activities and behaviors, to predict behavioral deterioration. We explore deterioration patterns from consecutive combinations of behavior classes corresponding to misbehavior, utilizing two publicly available datasets. We achieve F-1 scores as high as 0.7 with the initially-built model and 0.6 when alternative features are added to the initially-built model. Our method provides a straightforward way to obtain signals relevant to deterioration without involving other contributing factors, such as an individual’s mental state, personality traits, and affinity relationships [186]. Some of these opportunities are discussed in Section 7.5; i.e., fixing deterioration threshold and building a holistic model for determining the magnitude of deterioration.

This problem leaves room for future research. In the future, we aim to add multi-modal analysis and investigate behavioral sequences without converting misbehavior-related classes into a single class category. Furthermore, we would like to work on measuring the distance and similarity between multiple behavioral sequences¹², pre-

12. To this end, we plan to address the behavioral sequences as biological sequences in order to apply sequence alignment-based algorithms such as Needleman–Wunsch [136], Smith–Waterman [171], etc.

CHAPTER 7. ON PREDICTING BEHAVIORAL DETERIORATION IN ONLINE DISCUSSION FORUMS

dicting affinity relationships between individuals who exhibit deteriorating behaviors, identifying among these individuals those who seem to foment misbehavior within the online discussion forums, and assessing the likelihood that their affinity may evolve and the risks they may represent.

Chapter 8

Discovery of Temporal Deterioration Patterns from Behavioral Sequences

8.1 Introduction

This work is related to temporal patterns, misbehavior effects, psychological considerations, and behavioral deterioration prediction.

Temporal patterns. Recent works have addressed the problem of extracting temporal patterns from interval-based data by introducing novel machine learning models [30, 67, 130, 176, 207] and techniques for healthcare and psychology [3, 26, 91]. Temporal patterns can be explored for multiple purposes and in several different dimensions, depending highly on the nature of the problem. Yang and Leskovec [207], for instance, examined temporal patterns associated with online content and how the content’s popularity evolves and vanishes over time. Sultana and Gavrilova [176] studied a set of idiosyncratic features to automate identity verification using temporal profiles of users from social media. Jarmolowicz et al. [91] observed temporal patterns of behavior in quiz taking and found that students often allocate academic behavior in the form of a positive scallop when presented with externally assigned

deadlines. Our work here is different as we are not trying to find a unifying global model to examine popularity, identity verification, and procrastination, but rather propose techniques for detecting relevant temporal deterioration patterns within behavioral sequences exhibited by people at school, in prison, on social media and in communities in real-life.

Misbehavior effects. Research has yielded alarming results concerning the surge of misbehavior in terms of cyberbullying [18, 118] and psychological and economic victimization [5] in places such as schools [61, 89, 139, 177, 179], prisons [132, 182], organizations [192] and social media [31].

More generally, schools and jails follow closely disciplinary problems. While the study of [89] investigated peer misbehavior effects in the classroom, Sun and Shek [179] collected a video dataset for examining students’ behaviors in classroom scenes; identified the most common disruptive and unacceptable student problem behaviors such as disrespecting teaching and verbal aggression; and found that some behaviors may escalate in terms of frequency and intensity and can be contagious to some extent. Similarly to the preceding work, the study of [5] utilized closed-circuit television (CCTV) in prisons to study its effects on prison misbehavior and indicated that violent and unplanned misbehavior was relatively more likely to occur in view of CCTV coverage than was nonviolent and planned misbehavior; and showed that psychological and economic victimization¹³ occur frequently in prison and their consequences are widespread and potentially have a negative impact on inmates, correctional staff and organizations, public safety, etc. In order to identify the root cause of misbehavior and individuals who seem to foment misbehavior in a community, Morris et al. [132] examined how environmental strain measured at the prison-level influences inmate’s violent misconduct; constructed a model counting violent misconduct for investigating the trajectory of each inmate and assessed whether the strain of the environment distinguishes between trajectories. In this light, our work seeks to examine the persistent accumulation of misbehavior at the community level and to extract features from this persistence to verify whether accumulated misbehavior patterns from

13. – Psychological victimization (inmate-on-inmate verbal abuse or threats and inmate-on-officer verbal abuse or threats) – Economic victimization (theft, extortion, and robbery)

8.1. Introduction

the trajectories of certain community members have an influence on others. The rationale for doing so is to preemptively discover behavior signals that can encourage and foment misbehavior. It is important to note that persistent accumulation of misbehavior may have effects of infringing norms and expectations and major and devastating consequences on social harmony.

Psychological considerations. Recently, studies have examined psychological factors that could be engendered by misbehavior. More specifically, research showed that some victims of misbehavior are more likely to self-harm and engage in suicidal behaviors [94], and experience some unpleasant consequences, including psychological and anxiety disorders [40, 87, 120, 209] and low self-esteem [50]; others even commit suicide [79]. In order to prevent the occurrence and magnitude of psychological factors and consequences that misbehavior can engender, our work proposes to investigate misbehavior patterns at the community level to predict whether the persistent accumulation of misbehaviors tends to deteriorate or to become more harmful to the community’s quietude.

Gaps in the current deterioration studies. Previous computational work in behavioral deterioration detection focused on extracting individual-level characteristics related to behaviors exhibited by people for predicting deterioration from an online community [187]. The study showed that consecutive combinations of sequential patterns corresponding to misbehavior provide more reliable features for deterioration prediction. We observe some limitations of this study: one of them is the fact that prediction does not include the influence that the whole community may have on a deterioration in the behavior of certain community members. Relying solely on individual-level features to predict deterioration, in of itself, is not necessarily problematic, but this may render a significant proportion of deterioration patterns an untapped resource of potential. We believe that the aforementioned limitation can significantly hamper efforts to recognize and explore the objective factors underlying behavioral deterioration.

In this work, we seek to further understand the underlying temporal deterioration patterns that influence behavioral deterioration at the individual level, but with an

emphasis on the role, that behavioral deterioration at the community level may play in individual effects.

8.2 Method

In this section, we detail the formal notations and definitions used in our previous work (see §7).

Notations. Let $C = \{c_1, c_2, \dots, c_H\}$ denote the set of community members, and $\beta_B^C = \{B_1, B_2, \dots, B_L\}$ be the set of behavioral sequences exhibited by a community C , where $B_s = \{b_{t_1}, b_{t_2}, \dots, b_{t_T}\}$ and $s \in \{1, \dots, L\}$. Each b_{t_k} represents a behavior class exhibited by a c_i at time t_k , where $\forall k \in \{1, \dots, T\}$, $i \in \{1, \dots, H\}$ and $b_{t_k} \in \{N, M\}$. The classes N and M designate normal behavior and misbehavior, respectively. The sequence B_s denotes the concatenation of all behavior classes b_{t_k} exhibited by c_i over time. To better illustrate the behavioral sequence, Figure 8.1 shows the trajectory of behaviors exhibited by a community member c_i over time. To predict behavioral deterioration from β_B^C , we aim to extract features from behavioral sequences to capture signals that are potentially relevant to deterioration.

Definition 7 (Behavioral deterioration) *We define behavioral deterioration as the accumulation of misbehavior within a behavioral trajectory.*

For instance, let $B_1 = \text{NMMMMNMMMMNM}$ represent a behavioral sequence; the subsequences MM, MMM and MMMM denote the accumulation of misbehavior. We assume that these patterns can be considered as important deterioration-relevant signals.

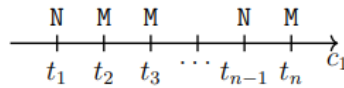


Figure 8.1 – Illustration of a behavioral sequence

Preliminary. In our work, we extract temporal deterioration patterns at the community level that have an influence on behavioral deterioration at the individual level.

8.2. Method

To this end, we structurally align behavioral sequences exhibited by all community members over time (see Table 8.2a). For the sake of simplicity, we refer to this alignment in this work as a behavioral matrix (BM). We seek to explore the behavioral matrix horizontally and vertically to capture deterioration patterns that may play in individual effects or may have an influence on individual trajectories. Note that the order of elements in behavioral sequences respects the timestep in which each behavior class was exhibited. Each column of the behavioral matrix represents the behavior class exhibited by each community member c_i at a specific timestep t_k (see Table 8.2b).

To predict behavioral deterioration, we examine the BM at the individual and community level. Specifically, at the individual level, we horizontally extract some statistically over-represented patterns across behavioral sequences and then investigate whether these patterns are deterioration-relevant signals. At the community level, we vertically extract features from the columns of the behavioral matrix to discover conserved deterioration patterns that can be observed when analyzing behavioral sequences horizontally and that may play a pivotal role in and eventually contribute to individual behavioral deterioration.

We believe that conserved deterioration patterns can provide important clues to thoroughly investigate whether the community has an influence on individual trajectories (see Table 8.2b). These clues outline broad trends of behavioral deterioration in a particular timestep. For instance, at time t_1 , by referring to feature extraction strategies proposed by [187], we observe that the patterns MM and MMM occur an excessive number of times in the first column of the behavioral matrix and are respectively highly conserved. We notice that these patterns are more highly conserved in the remainder of c_1 and c_5 's behavioral sequences from t_2 . In Section 8.2.1, we briefly describe the feature extraction process utilized to discover conserved deterioration patterns in the columns of the behavioral matrix.

8.2.1 Feature extraction

Tshimula et al. [187] proposed feature extraction strategies to analyze behavioral sequences. They designed character n-gram features to capture signals that are po-

CHAPTER 8. Discovery of Temporal Deterioration Patterns from Behavioral Sequences

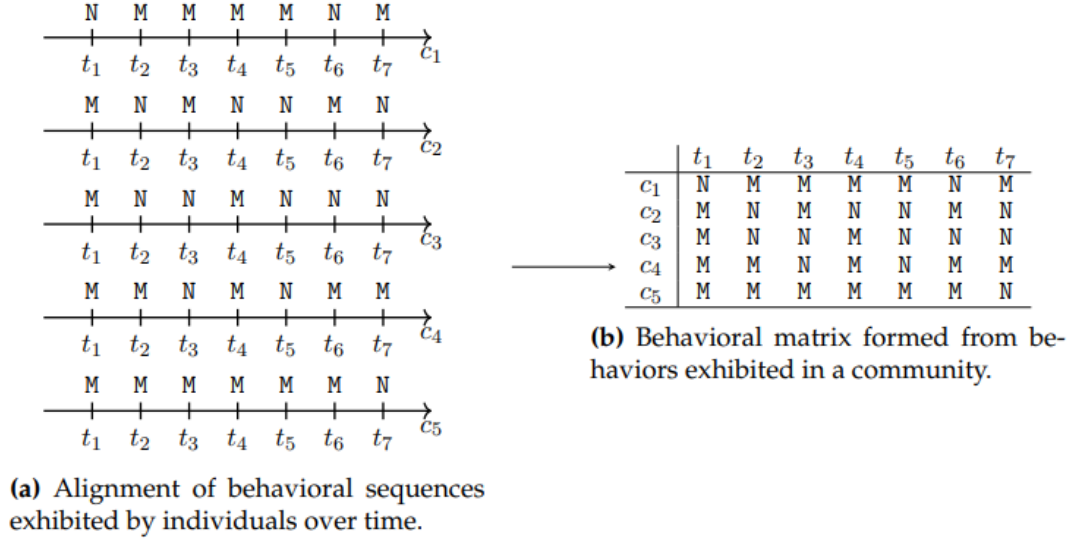


Figure 8.2 – Illustration of the alignment of community behavioral trajectories in a behavioral matrix

tentially relevant to deterioration and extracted n-gram features with a pair of values (h_p, v_p) ; where h_p denotes the n-gram feature p and v_p represents the count of the feature within behavioral sequences.

The n-grams can be engendered by sliding a window of length n over the B_s or the sequence formed from a BM's column. Figure 7.1 depicts how n-gram features can be extracted from a B_s . More specifically, the features extracted from the B_s illustrated in Figure 7.1 can be presented as follows: 2-grams $\{(NM,3), (MN,2), (MM,5)\}$, 3-grams $\{(MMM,3), (NMM,2), (MMN,2), (MNM,2)\}$, 4-grams $\{(MMMM,2), (NMMM,2), (MMNM,2), (MMMM,1), (MNMM,1)\}$ and 5-grams $\{(MMMMM,2), (NMMMM,1), (MMMMN,1), (MMNMM,1), (MNMMM,1), (NMMMM,1)\}$.

This feature process maps n-gram that support the accumulation of misbehavior classes as *deterioration* and other n-grams as *non-deterioration*, and differently treats the features for which the length is greater than or equal to 4 (e.g., 4-grams and 5-grams); especially, features which do not fully support the accumulation of misbehavior as full-fledged behavioral sequences (e.g., NMMM and NMMMMN). To assign corresponding deterioration labels in such cases, sub-2-grams are examined within these features by applying the same logic as in Figure 7.1(a). The main reason for

8.2. Method

investigating sub-2-grams is to better track the momentum of the accumulation of different behavior classes and discover deterioration patterns. The n -gram features ($n \geq 4$) are labeled as *deterioration* based on whether the majority of the sub-2-grams they contain support the accumulation of misbehavior classes. For instance, NMMM, $\{(NM,1), (MM,2)\}$; and NMMMN comprises $\{(NM,1), (MM,2), (MN,1)\}$. Specifically, NMMM is labeled as *deterioration* and NMMMN as *non-deterioration*.

8.2.2 Predictive models

We utilize advanced variants of Recurrent Neural Networks to predict temporal behavioral deterioration. The advantage of utilizing these variants is that they efficiently deal with long term dependencies within sequences such as behavioral sequences.

Bidirectional long-short term memory (BiLSTM). Recurrent Neural Networks (RNNs) are a class of neural networks that allow previous outputs to be used as inputs while having hidden states. It simply means that RNNs have a memory that stores and captures information about what has been calculated so far and passes that memory to the next node. This allows it to exhibit temporal dynamic behavior, although they suffer from short-term memory when they deal with large sequential data. The vanishing and exploding gradient phenomena are often encountered in the context of RNNs. Basically, gradients are values used to update the weights of a neural network. The vanishing gradient problem is when the gradient shrinks as it back propagates through time. If a gradient value becomes extremely small, it does not contribute too much learning. Exploding gradients are a problem where large error gradients accumulate and result in very large updates to neural network model weights during training. These problems justify the reason why RNNs have difficulty capturing long-term dependencies.

To this end, many RNN variants have been introduced: GRU (gated recurrent units) [36], (long-short term memory) LSTM [80], bidirectional LSTM [62], attention-based model [193], etc. These methods propose effective solutions for mitigating short-term memory problems encountered by traditional RNNs and for enabling neural

networks to capture much longer range dependencies using mechanisms called gates. Gates are neural networks that regulate the flow of information circulating through the sequence chain. LSTM incorporates several control gates and a constant memory cell, the details of which are following (we borrowed the description from [205]):

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \quad (8.1)$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \quad (8.2)$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \quad (8.3)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \phi(W_{cx}x_t + W_{ch}h_{t-1}) \quad (8.4)$$

$$h_t = o_t \odot \phi(c_t) \quad (8.5)$$

Each unit receives an input h_{t-1} from its previous unit together with the input x_t at the time point t . Each unit has its memory updating the previous memory c_{t-1} with the current input modulation. The network takes three inputs: x_t , h_{t-1} , and c_{t-1} , and has two outputs: h_t (the output of the current cell state) and c_t (the current cell state). Three gates are separately utilized to control input (Eq. 8.1), forget (Eq. 8.2) and output (Eq. 8.3). More specifically, the input gate i_t controls how much influence the inputs x_t and h_{t-1} exerts to the current memory cell (Eq. 8.1). The forget gate f_t controls how much influence the previous memory cell c_{t-1} exerts to the current memory cell c_t (Eq. 8.2). Output gate controls how much influence the current cell c_t has on the hidden state cell h_t (Eq. 8.3). The memory cell unit c_t is a summation of two components: the previous memory cell unit c_{t-1} , which is modulated by f_t and $\phi(W_{cx}x_t + W_{ch}h_{t-1})$, and a weighted combination of the current input and the previous hidden state, modulated by the input gate i_t (Eq. 8.4). In addition, cell state is filtered with the output gate o_t for a hidden state updating (Eq. 8.5), which is the final output from an LSTM cell.

Referring to behavioral sequences formulated in §8.2, W_x matrices (containing weights applied to the current input in form of feature as extracted in §8.2.1) and W_h matrices (representing weights applied to the previous hidden state) can be learned, the vector b_i , b_f and b_o are biases for each layer, σ and ϕ denote the sigmoid and hyperbolic non-linear functions, and \odot indicates element-wise multiplication operation.

8.2. Method

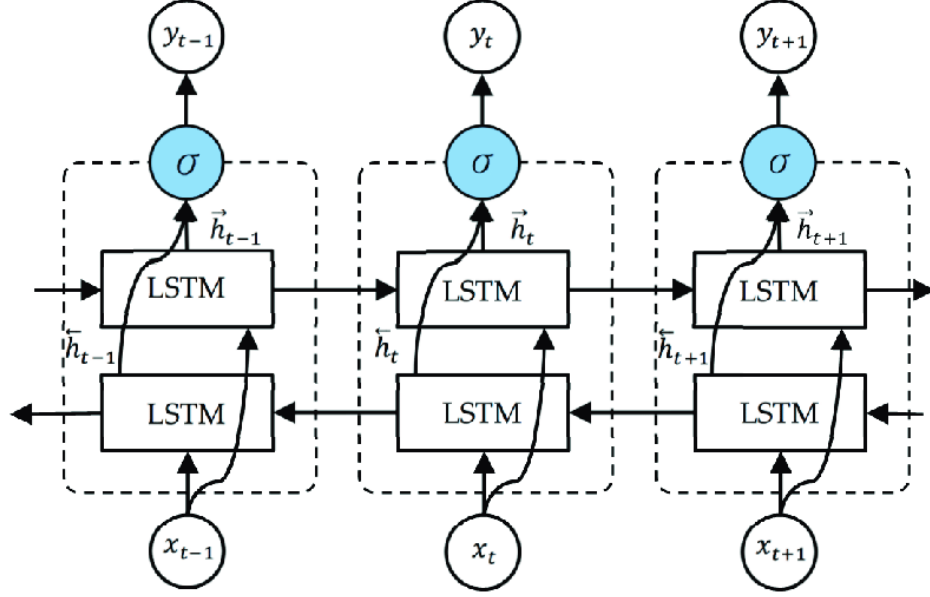


Figure 8.3 – Overview of the bidirectional LSTM model (source: <https://tinyurl.com/3s7byvzh>)

Bidirectional LSTM (BiLSTM). The network structure of BiLSTM can be roughly seen as two layers: forward LSTM layer and backward LSTM layer. The network structure of BiLSTM is shown in Figure 8.3. We denote a process of an LSTM cell as H . The general operation process of BiLSTM is as follows: at time t , it first passes through each LSTM unit in the forward layer and then forward calculates and saves the backward hidden layer output: \vec{h}_t at each time. After passing through each LSTM unit in the backward layer, the output of the forward hidden layer at each moment is calculated in reverse: \overleftarrow{h}_t [202]. Finally, \vec{h}_t and \overleftarrow{h}_t are synthesized at each moment to the final output y_t . The W_x and W_h matrices in Eq. 8.6 and 8.7 are the same as those in Eq. 8.1, 8.2 and 8.3. The $W_{\vec{h}_y}$ (representing weights applied to the forward hidden state) and $W_{\overleftarrow{h}_y}$ (representing weights applied to the backward hidden state) are learned with behavioral sequences.

$$\vec{h}_t = H(W_{x\vec{h}}x_t + W_{h\vec{h}}\vec{h}_{t-1} + \vec{b}_h) \quad (8.6)$$

$$\overleftarrow{h}_t = H(W_{x\overleftarrow{h}}x_t + W_{h\overleftarrow{h}}\overleftarrow{h}_{t-1} + \overleftarrow{b}_h) \quad (8.7)$$

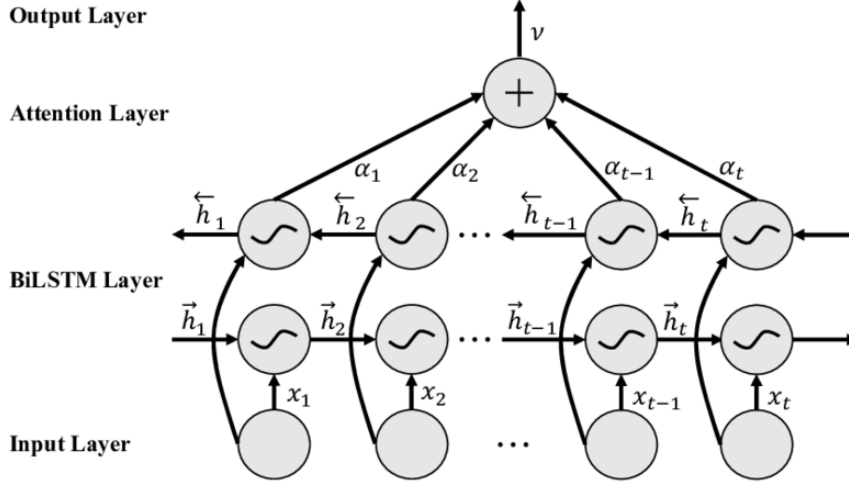


Figure 8.4 – Architecture of BiLSTM with attention (Source: <https://tinyurl.com/37kt4857>)

$$y_t = W_{\vec{h}_y} \vec{h}_t + W_{\overleftarrow{h}_y} \overleftarrow{h}_t + b_y \quad (8.8)$$

BiLSTM with attention mechanism (BiLSTM + Attention). Attention mechanism aims to show the potential of automatically dismissing unnecessary information in the entire input sequence and highlighting the relevant parts. The main idea is to induce attention weights over the input sequence to prioritize the set of positions where relevant information is present for generating the next output token [27, 193]. Let $\vec{h}_1, \vec{h}_2, \dots, \vec{h}_t$ and let $\overleftarrow{h}_1, \overleftarrow{h}_2, \dots, \overleftarrow{h}_t$ denote respectively the output of the LSTM forward layer and the LSTM backward layer for a behavioral sequence of length t . We define $\vec{\beta}_i$ by utilizing \vec{h}_t to attend to all its previous outputs $\vec{h}_1, \vec{h}_2, \dots, \vec{h}_{t-1}$, and $\overleftarrow{\beta}_i$ by utilizing \overleftarrow{h}_t to attend to all its previous outputs $\overleftarrow{h}_1, \overleftarrow{h}_2, \dots, \overleftarrow{h}_{t-1}$:

$$\vec{\beta}_i = \frac{\exp(\phi(W [\vec{h}_t; \vec{h}_i]))}{\sum_{i=1}^{t-1} \exp(\phi(W [\vec{h}_t; \vec{h}_i]))}, \quad (8.9)$$

$$\overleftarrow{\beta}_i = \frac{\exp(\phi(W [\overleftarrow{h}_t; \overleftarrow{h}_i]))}{\sum_{i=1}^{t-1} \exp(\phi(W [\overleftarrow{h}_t; \overleftarrow{h}_i]))}, \quad (8.10)$$

where ϕ denotes \tanh . The final representation for the behavioral sequence is obtained

8.3. Deterioration prediction

Table 8.1 – Results of the individual level prediction of behavioral deterioration. Bold font indicates the best results for each class label.

	Class	HatebaseTwitter	TRAC
Ours+SVM (see Table 7.1)	Deterioration	0.722	0.785
	Non-deterioration	0.718	0.749
Ours+LR (see Table 7.1)	Deterioration	0.72	0.761
	Non-deterioration	0.719	0.758
Baseline (LSTM) (see Table 7.1)	Deterioration	0.719	0.737
	Non-deterioration	0.716	0.733
BiLSTM	Deterioration	0.725	0.81
	Non-deterioration	0.727	0.813
BiLSTM+Attention	Deterioration	0.751	0.818
	Non-deterioration	0.747	0.814

by weighted summation of value as follows:

$$\vec{a} = \sum_{i=1}^{t-1} \vec{\beta}_i \cdot \vec{h}_i, \quad (8.11)$$

$$\overleftarrow{a} = \sum_{i=1}^{t-1} \overleftarrow{\beta}_i \cdot \overleftarrow{h}_i, \quad (8.12)$$

$$\vec{v} = [\vec{a}; \vec{h}_t], \quad (8.13)$$

$$\overleftarrow{v} = [\overleftarrow{a}; \overleftarrow{h}_t], \quad (8.14)$$

where \vec{a} and \overleftarrow{a} are respectively the weighted sum of all previous outputs in the LSTM forward and backward layer; \vec{v} and \overleftarrow{v} are the final representation for the behavioral sequence, which are respectively concatenated by \vec{a} and \overleftarrow{a} and the last output \vec{h}_t and \overleftarrow{h}_t . Specifically, we take the forward representation \vec{v} and backward representation \overleftarrow{v} and then concatenate them and feed the resulting vector to a standard softmax layer for classification (see Figure 8.4).

8.3 Deterioration prediction

We utilized the same datasets, experimental settings and model evaluation with §7 as described in §7.4. We predict behavioral deterioration using feature extracted

Table 8.2 – Results of the community level prediction of behavioral deterioration. Bold font indicates the best results for each class label.

	Class	HatebaseTwitter	TRAC
Ours+SVM (see Table 7.1)	Deterioration	0.751	0.8
	Non-deterioration	0.735	0.776
Ours+LR (see Table 7.1)	Deterioration	0.748	0.783
	Non-deterioration	0.766	0.812
Baseline (LSTM) (see Table 7.1)	Deterioration	0.724	0.757
	Non-deterioration	0.731	0.758
BiLSTM	Deterioration	0.773	0.83
	Non-deterioration	0.777	0.835
BiLSTM+Attention	Deterioration	0.803	0.828
	Non-deterioration	0.8	0.839

at the individual and community level from the behavioral matrix.

Tables 8.1 and 8.2 show the performance of our models at individual and community level. We report that the novel method proposed in §8.2 shows the ability to predict behavioral deterioration, with F-1 scores of over 0.8 for both predicted labels and yields statistically significant improvements over our previous results, surpassing the previous models by 0.029 for *Deterioration* and 0.028 for *Non-deterioration* on HatebaseTwitter and 0.033 for *Deterioration* and 0.056 for *Non-deterioration* on TRAC at the individual level. One of the advantages of the novel models over our previous models is its ability of predicting deterioration at the community level. It can be seen that the novel models predicted behavioral deterioration with high accuracy at the community level, supporting our theoretical assumptions that analyzing deterioration at the community level can lead to the discovery of potential deterioration patterns. These results suggest that performing the analysis of bidirectional relationships between deterioration classes within behavioral sequences can lead to significantly better performance, in particular when deterioration labels are well defined. BiLSTM-models performed very well. Specifically, while we combine bidirectional LSTM with attention mechanism (BiLSTM+Att), our novel models performed more accurately and consistently than our previous models. We note that even our second best-performing classifier attained better performance on experimental datasets than our previous models.

8.3. Deterioration prediction

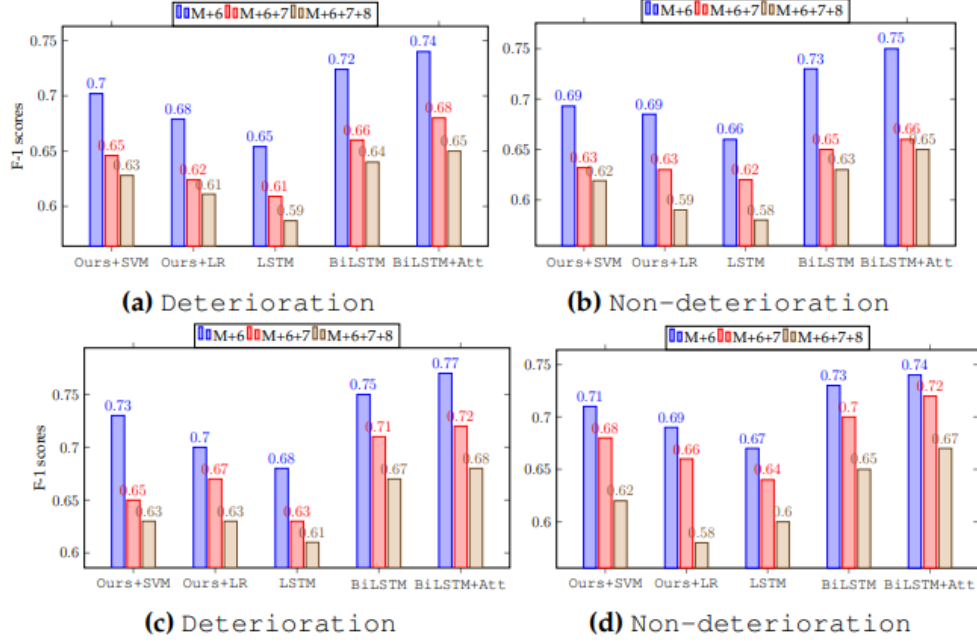


Figure 8.5 – Results of the individual level prediction of behavioral deterioration by adding extra features to the main model. Specifically, M+6 means that we add 6-grams to the model, M+6+7 stands for 6- and 7-grams and M+6+7+8 means that we add 6-, 7- and 8-grams. We predict (a) and (b) on HatebaseTwitter and (c) and (d) on TRAC. Note that BiLSTM+Att indicates BiLSTM with attention.

To examine the effect the number of features has on model performance, we extend the initially-built model by including in it some additional features to investigate deterioration patterns within behavioral sequences. Figures 8.5 and 8.6 present the results of behavioral deterioration prediction with additional features at the individual and community level. At the individual level, we note that the best performances yielded by our models exceed 0.65; that is, (0.742, 0.68, 0.653) for *Deterioration* and (0.75, 0.661, 0.65) for *Non-deterioration* with M+6, M+6+7, and M+6+7+8, respectively, on HatebaseTwitter; and (0.773, 0.724, 0.68) for *Deterioration* and (0.74, 0.722, 0.673) for *Non-deterioration* with M+6, M+6+7, and M+6+7+8, respectively, on TRAC. At the community level, we report that the best performances yielded by our models exceed 0.7; that is, (0.763, 0.709, 0.724) for *Deterioration* and (0.772,

CHAPTER 8. Discovery of Temporal Deterioration Patterns from Behavioral Sequences

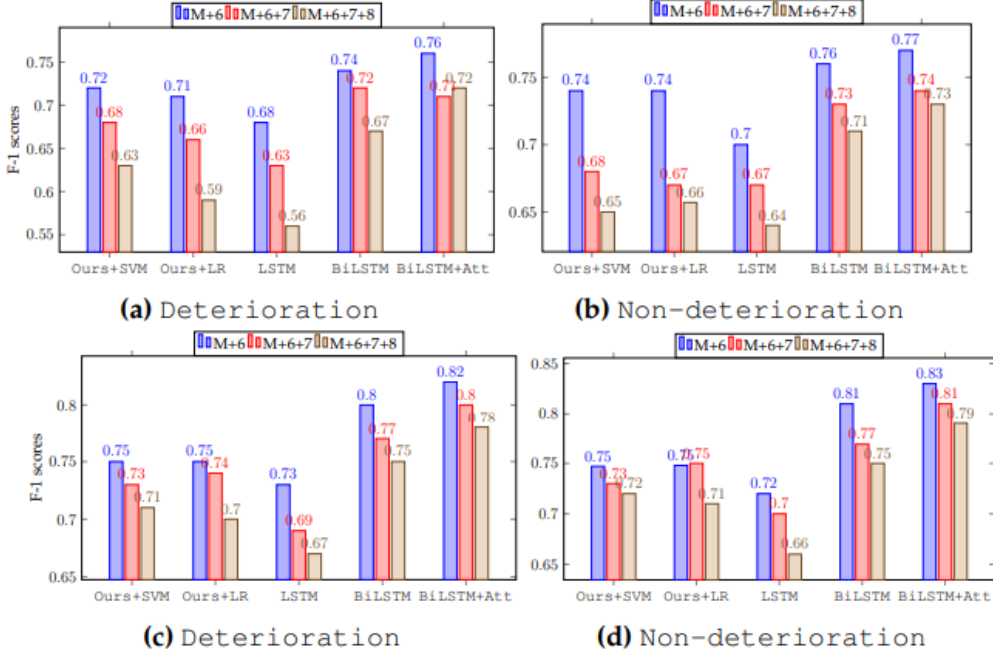


Figure 8.6 – Results of the community level prediction of behavioral deterioration by adding extra features to the main model.

0.74, 0.733) for *Non-deterioration* with M+6, M+6+7, and M+6+7+8, respectively, on HatebaseTwitter; and (0.818, 0.8, 0.784) for *Deterioration* and (0.83, 0.811, 0.793) for *Non-deterioration* with M+6, M+6+7, and M+6+7+8, respectively, on TRAC. These results support the observation made in §7, i.e., the model performance decreases when the number of features increases. Surprisingly, we observe that prediction performance at the community level for supplementary features largely surpasses the performance achieved by our previous work on the initially-built models.

Importantly, our models do not implicitly include emotional states that could potentially lead individuals to express persistent accumulations of misbehavior. The combination of emotional states and our features in a single model can make it difficult to extract signals relevant to deterioration, because there are various emotional states. We propose to examine separately the emotional states from social interactions associated with the extracted features. To gain a better understanding of the differences between emotional across behavioral sequences, we investigate text

8.4. EMOTIONAL STATES AND DETERIORATION

Table 8.3 – Prediction quality for emotional states at the individual level, as measured using the Pearson r . Note that 2-grams indicates MM; 3-grams, MMM; 4-grams, (MMMM, NMMM, etc.) and 5-grams, (MMMMM, NMMMM, etc.); and these results concern sub-datasets indicating the presence of deterioration patterns ($*p \leq 0.05$).

	2-grams	3-grams	4-grams	5-grams		2-grams	3-grams	4-grams	5-grams
Anxiety	0.34	0.165	0.281	0.215	Anxiety	0.253	0.22	0.285	0.316
Stress	0.33	0.33	0.221	0.329	Stress	0.272	0.289	0.347	0.408
Fear	0.191	0.207	0.322	0.16*	Fear	0.195	0.16	0.295	0.33*
Anger	0.408*	0.374*	0.312*	0.29*	Anger	0.403*	0.353*	0.3*	0.298*
Sadness	0.223	0.178	0.181	0.277	Sadness	0.13	0.127	0.207	0.257
Disgust	0.147	0.149	0.204	0.261*	Disgust	0.173	0.112	0.186	0.204
Surprise	0.16	0.182	0.238*	0.291*	Surprise	0.153	0.156	0.179	0.19

(a) HatebaseTwitter
(b) TRAC

Table 8.4 – Prediction quality for emotional states at the community level, as measured using the Pearson r . Note that 2-grams indicates MM; 3-grams, MMM; 4-grams, (MMMM, NMMM, etc.) and 5-grams, (MMMMM, NMMMM, etc.); and these results concern sub-datasets indicating the presence of deterioration patterns ($*p \leq 0.05$).

	2-grams	3-grams	4-grams	5-grams		2-grams	3-grams	4-grams	5-grams
Anxiety	0.336	0.279	0.273	0.385	Anxiety	0.203	0.26	0.262	0.279
Stress	0.165	0.218	0.158	0.258	Stress	0.138	0.121	0.207	0.246
Fear	0.181	0.19	0.253*	0.347*	Fear	0.205	0.266	0.33*	0.335*
Anger	0.412*	0.399*	0.224*	0.195*	Anger	0.378*	0.367*	0.269*	0.226*
Sadness	0.226	0.301	0.307	0.413	Sadness	0.196	0.283	0.351	0.368
Disgust	0.285	0.263	0.351	0.388*	Disgust	0.276	0.273	0.316	0.407
Surprise	0.219	0.228	0.312*	0.394*	Surprise	0.282	0.24	0.319*	0.349*

(a) HatebaseTwitter
(b) TRAC

data to discover emotional states that individuals manifest and assess whether these emotional states progressively contribute to the degradation of individual behaviors.

8.4 Emotional states and deterioration

An individual's behavior may undergo a sudden or gradual deterioration, and this phenomenon may happen under the influence of several factors, including friends, self-commitment, and personality issues. Personality can influence judgments and decisions in various ways [113, 213]. Personality and emotion regulation are related

CHAPTER 8. Discovery of Temporal Deterioration Patterns from Behavioral Sequences

Table 8.5 – Prediction quality for emotional states at the individual level, as measured using the Pearson r . Note that these results concern sub-datasets indicating the absence of deterioration patterns ($*p \leq 0.05$).

	2-grams	3-grams	4-grams	5-grams		2-grams	3-grams	4-grams	5-grams
Anxiety	0.404*	0.395*	0.338*	0.272	Anxiety	0.444*	0.415*	0.364*	0.263
Stress	0.399*	0.343*	0.274	0.288	Stress	0.37*	0.362*	0.323*	0.293
Fear	0.361*	0.352*	0.341*	0.405*	Fear	0.334*	0.333*	0.328*	0.262*
Anger	0.149	0.159	0.26	0.281*	Anger	0.158	0.158	0.22	0.226*
Sadness	0.235	0.287*	0.325*	0.342*	Sadness	0.36*	0.345	0.399*	0.411*
Disgust	0.285*	0.294*	0.397*	0.41*	Disgust	0.218	0.261	0.347*	0.389*
Surprise	0.37*	0.331*	0.305*	0.168	Surprise	0.32*	0.319*	0.304*	0.197

(a) HatebaseTwitter
(b) TRAC

Table 8.6 – Prediction quality for emotional states at the community level, as measured using the Pearson r . Note that these results concern sub-datasets indicating the absence of deterioration patterns ($*p \leq 0.05$).

	2-grams	3-grams	4-grams	5-grams		2-grams	3-grams	4-grams	5-grams
Anxiety	0.435*	0.433*	0.346*	0.381*	Anxiety	0.413*	0.383*	0.272	0.229*
Stress	0.391*	0.353*	0.279	0.263	Stress	0.376*	0.364*	0.272*	0.225
Fear	0.34*	0.352*	0.392*	0.418*	Fear	0.341*	0.374*	0.359*	0.417*
Anger	0.151	0.154	0.169	0.182	Anger	0.161	0.168	0.171	0.177
Sadness	0.243	0.249*	0.348*	0.407*	Sadness	0.395*	0.332*	0.36*	0.408*
Disgust	0.261	0.308*	0.299*	0.342*	Disgust	0.214	0.319*	0.334*	0.396*
Surprise	0.379*	0.367*	0.296*	0.159	Surprise	0.396*	0.328*	0.316*	0.305*

(a) HatebaseTwitter
(b) TRAC

but distinct and both contribute to providing a comprehensive description of human affective experience [95]. In this research, we examine the trajectory of behavioral deterioration in order to discover emotional states that individuals gradually display and evaluate whether it contributes towards dramatically worsening the behavior over time. We analyze emotional states that have a greater relationship with deterioration. We scrutinize emotional states that may affect personality change, including *anxiety*, *stress*, *fear*, *anger*, *sadness*, *disgust*, and *surprise*.

To this end, we created sub-datasets by assembling the features extracted in §8.2.1 based on their lengths and on the labels in which they belong. For instance, we put the 2-gram features supporting the accumulation of misbehavior classes in a single sub-dataset and all 2-gram features indicating the absence of deterioration signals (such as NM and MN) in a different sub-dataset. We applied this logic to other features by

8.4. EMOTIONAL STATES AND DETERIORATION

considering their size and the label signals they indicate. For each sub-dataset, we took all social interactions (text data) associated with each class that composes the features.

To measure emotional state exhibited [57] in text data, we utilize Linguistic Inquiry and Word Count (LIWC) [144], a dictionary which is widely employed in computational linguistics as a source of features for psychological and psycholinguistic analysis. LIWC comprises words that have very clear, pre-labeled meanings. The dictionary includes words in various categories, notably linguistic dimensions, psychological processes and personal concerns. Each category is found to be correlated with several psychological traits and outcomes [65, 66]. Specifically, we focus on the psychological processes category in order to explore the linguistic usage in text data. We leverage each social interaction in text data and measure the proportion of word tokens that fall into *anxiety*, *stress*, *fear*, *anger*, *sadness*, *disgust* and *surprise*.

To predict emotional states in text data, we treat each sub-dataset separately and stratify each sub-dataset by 10-fold cross-validation to split our training and test sets. Linear regression with elastic net regularization was performed to predict the emotional state signals derived from the LIWC characteristics and to evaluate the quality of the prediction using the Pearson correlation (r) as an evaluation measure. Tables 8.3, 8.4, 8.5 and 8.6 show the quality of prediction of emotional states on an individual and community level in behavioral sequences. We report that Pearson correlation coefficients for *anger* are all statistically significant ($p < 0.05$) for sub-datasets indicating the presence of deterioration patterns at the individual and community level (Tables 8.3 and 8.4). We observe that emotional states such as *anxiety*, *stress*, *fear* and *disgust* are not statistically significant in sub-datasets indicating the presence of deterioration patterns, except for *surprise* ($p < 0.05$) in Tables 8.3(a), 8.4(a) and 8.4(b). These results show that individuals perpetrating deteriorating behavior manifest an important proportion of *anger* and do not display *anxiety*, *stress*, *fear* and *disgust* while exhibiting their misdeeds. Evidence shows that *anger* could be the principal emotional state that contributes considerably to the degradation of behaviors. Particularly, we observe that Pearson’s correlations are statistically significant for *disgust* in 5-grams (Tables 8.3(a), 8.4(a) and 8.4(b)), and for *fear* in 5-grams (Table 8.3) and in 4- and 5-grams (Table 8.4).

We note that Pearson’s correlations for *anxiety*, *fear*, *surprise* and *sadness* are statistically significant ($p < 0.05$) in sub-datasets indicating the absence of deterioration patterns at the individual and community level (Tables 8.5 and 8.6). Particularly, we observe that Pearson’s correlations for *anger* are not statistically significant ($p > 0.05$); and also note that Pearson’s correlations for *disgust* are statistically significant ($p < 0.05$) for the majority of sub-datasets. These results show that sub-datasets indicating the absence of deterioration patterns do not contain a significant proportion of *anger* and exhibit the feelings such as *anxiety*, *fear*, *surprise* and the expression of *disgust* at the individual and community level.

8.5 Discussion and conclusion

Through this work, we have introduced a methodology to help identify relevant signals in behavioral sequences that have a greater potential of transitioning to deterioration. We proposed models that predict behavioral deterioration at the individual and community level. An important contribution of our methodology, in particular, has been the ability to identify highly conserved patterns which indicate ripe signals for investigating deterioration at the community level. There are clear benefits to using our methodology, as demonstrated by two bidirectional LSTMs – this provides strong signals relevant to deterioration, and some intuitive and interpretable groupings of features without significant manual intervention. We show the ability of our models in predicting behavioral deterioration with a high degree of accuracy, i.e., F-1 scores of over 0.8. We also demonstrate the robustness and effectiveness of our models by extending the number of features. Our results yielded statistical improvements over our previous models and our findings suggest that the analysis of bidirectional relationships between deterioration classes within behavioral trajectories can lead to significantly better performance.

Furthermore, we investigate the trajectory of behavioral deterioration in order to discover the emotional states that individuals gradually manifest and evaluate whether these contribute to the degradation of behaviors over time. More importantly, we show the utility of examining emotional states separately. By examining correlations between emotional states and the various features extracted, we provide

8.5. DISCUSSION AND CONCLUSION

some insight that indicates the effects of deteriorating signals on emotional states, and vice versa; our findings suggest that *anger* could be a potential emotional state that can substantially contribute to behavioral deterioration.

Crucially, we believe that our models can pave the way for the prediction of behavioral deterioration using data from various sources such as jails, schools, and addiction treatment centers. For instance, our models and findings can be leveraged as a complementary screening tool and used in conjunction with ground-truth to gauge whether the behaviors of intended individuals deteriorate or improve. This work can help create provisions for early detection of misbehavior to deterioration.

While the results reported here hold promise for future work, both theoretical and applied, our research is limited by several important factors. All these experiments, taken together, indicate that there are a diverse set of quantifiable signals relevant to deterioration. They indicate that individual and community level analyses can be made more accurately and efficiently than previous methods, yet there remains as-of-yet untapped information. For instance, we rely heavily on persistent accumulations of misbehavior to predict behavioral deterioration. This could make it difficult to detect suddenly deteriorating behavior. Further study should propose a definition that includes several deterioration scenarios to discover untapped information.

Our results identify several opportunities and challenges in the development of pre-trained models that could understand linguistic features underlying the deterioration task and fine-tune this task based on the nature of data sources.

Conclusion

This chapter concludes this dissertation with a summary of the contributions and highlights some future directions for continued research.

Summary of This Dissertation

Social media platforms assemble individuals who have diversified convictions and beliefs to interact in friendly and civilized ways. Increasingly, however, they are having the opposite effect, due to a rising tide of deviations, and deliberate provocations; since some individuals engage in misbehavior that harms and adversely affects the equanimity of other users. Persistent accumulations of misbehavior could be a valid predictor of risk factors for behavioral deterioration. Early detection of behavioral deterioration can be of crucial importance in preventing individuals' misbehavior from escalating in severity. The problem of behavioral deterioration has not been extensively studied in the context of social media. In this dissertation, we divided this problem into three components (**affinity**, **personality**, and **deterioration**) to understand individuals exhibiting deteriorating behaviors and proposed machine learning models to investigate the underlying factors contributing to behavioral deterioration.

First, we deal with understanding emotional states and moral foundations in the language use of text data. The rationale behind this is to discover whether individuals perpetrating misbehavior manifest social morality and emotional instability. We investigate emotional states and social morality to understand moral differences in a broad spectrum of interactions on social media. Morality guides human social interactions and can potentially conduct to a divergence of opinion, polarity, and hostility when there is moral shock within a community. The key insight is to discover whether differences in moral dimensions have a certain influence on emotional states. To this

CONCLUSION

end, we build a machine learning model using a moral foundations dictionary and propose another model based on natural language inference to automatically extract morality features. We compute the Pearson correlation coefficients between morality features and psycholinguistic features extracted from text data to discover the influence of morality on emotions. Furthermore, we examine the temporal evolution of emotional states and identify relevant patterns that lead to emotional instability and breakdown in highly motivated high-conflict interactions such as law enforcement interviews. Through extensive experiments on four different datasets, our findings indicate emotional trajectories illustrating shifts in emotional states and show similarities and correlations between the emotional trajectories in efficient ways. In the task of predicting behavioral deterioration, the proposed approaches are crucial because they can help facilitate the understanding and reveal the involvement of emotional states in the language use of individuals for whom behaviors tend to deteriorate.

Next, we address the challenge of discovering affinity relationships between social media users. The problem of affinity relationships in the context of social media has not been clearly and formally defined in the literature. We propose mathematical definitions of affinity influence and extract several interpretable features from these definitions. The challenge of discovering affinity goes beyond carrying out an analysis based on structural features such as likes, shares, followers and followings. Rather than relying solely on structural features of social media, we combine structural features, temporal information and the content of interactions. We develop an advanced method based on Markov models, machine learning and natural language processing to quantify affinity scores using the combined features. We utilize the quantified affinity scores to investigate the evolution of affinity over time and predict affinity relationships arising from the influence of certain users. Through extensive experimental evaluation, we show that our approach achieves good performance on the experimental datasets, with an F-1 score of over 0.75 and statistically significant improvements over existing techniques, and results in robust discovery and considers minute details. In the task of predicting behavioral deterioration, the proposed approach can reveal individuals who seem to foment misbehavior on social media platforms and assess the likelihood that their relationships can evolve and the risks they may represent.

Furthermore, we investigate the influence of personality on affinity. The rationale

CONCLUSION

behind this is to discover affinity relationships between different personality types. The combination of affinity and personality allows us to understand how individuals with similar personality traits get to develop their affinity and discern what attracts an individual to another. In contrast to psychological research, we utilize the language use of text data to evaluate personality and emotional states ; we propose approaches that utilize psycholinguistic features, to measure emotional states and understand their linguistic idiosyncrasies. Specifically, we derive affinity relationships between individuals and examine personality based on the language use to discover the emotional stability of affinity relationships, and measure semantic similarity at the personality type level to understand the logic behind the development of affinity. The critical motivating insight is that our results identify influential personality types that weigh more heavily on affinity relationships and show that personality can be predicted from the spontaneous language with an F-1 score superior to 0.76. In addition, we find that semantic similarity and emotional (in)stability constitute an essential lead for understanding the implications of personality in the development of affinity. Our results identify several statistically significant correlations in terms of emotional stability in personality-based affinity relationships. Our outcomes' theoretical and practical implications can be valuable for supporting decision-making processes in various domains, including clinical psychology, forensic psychology, digital forensics, human factors and social science. In reality, investigations into the influence of personality on affinity can be driven by the concrete needs of applications ; for example, the role that personality plays in the effective functioning of behavioral deterioration.

Finally, we investigate the problem of behavioral deterioration and propose new models that construct behavioral sequences from temporal behaviors exhibited by individuals. Since this problem is relatively new in the context of social media, we study how the divergence of opinion can potentially conduct unhealthy conversations and emotional reactions and introduce a formal definition of the problem of behavioral deterioration. For stance classification, we construct a model on top of RoBERTa to classify stances by capturing the context of the discussion through the examination of pairs of stances and relational structures of discussion specific to each topic within the defined window of interactions of each participant of the discussion. We investigate the degree of disagreement and neutrality in the discussion to measure the divergence

CONCLUSION

of opinion on topics addressed in the discussion and predict the emotion associated with interactions by topic. For the prediction of behavioral deterioration, we propose new models to extract consecutive combinations of sequential patterns corresponding to misbehavior to predict behavioral deterioration at the individual level. We find that relying solely on individual-level features to predict deterioration, in of itself, is not necessarily problematic, but this may render a significant proportion of deterioration patterns an untapped resource of potential. Consequently, we investigate temporal deterioration patterns from behavioral sequences to predict deterioration at the community level.

Through extensive experiments on real-world datasets, our experiments suggest that our models have the potential of leveraging behavioral sequences for predicting signals relevant to deterioration from accumulations of behaviors and show the ability of our models in predicting behavioral deterioration with a high degree of accuracy, i.e., F-1 scores of over 0.8. Moreover, we investigate the trajectory of behavioral deterioration to discover the emotional states that individuals progressively manifest and evaluate whether these emotional states contribute to the deterioration of behaviors as time moves forward. Our results suggest that *anger* could be a potential emotional state that can substantially contribute to behavioral deterioration.

Each of these aforementioned components has its own scope in the context of behavioral deterioration and benefits the understanding of the transition from persistent accumulations of misbehavior to deterioration. We show the power of investigating these components for understanding the temporality of behavioral deterioration through their underlying psychological traits and emotional states and human relations. Our discoveries can be used by companies, schools, prisons, psychiatric centers and organizations for monitoring people manifesting signals relevant to behavioral deterioration; for instance, psychiatric centers can utilize our models to track the consecutive accumulation of daily signs of individuals with mental health conditions to predict signals relevant to deterioration or improvement. In prisons, our models can be used to predict the behavioral deterioration of recidivists and inmates stimulating defiant and aggressive behaviors. At schools, our models can be utilized as a barometer to measure behavior escalation and predict negative affinity relationships and behavioral deterioration from students breaking the behavior code and misbeha-

vors such as disruptive talking, chronic avoidance of work, clowning, interfering with teaching activities, harassing classmates, verbal insults, rudeness to teacher, defiance, hostility, absenteeism, bullying and other inappropriate behaviors. In companies, our models can be applied to predict the behavioral deterioration of employees engaged in code-of-conduct violations such as discrimination, gossiping, bad jokes, physical threats, negative remarks, and so on.

Our models can help the previously cited organizations anticipate actions and reinforce their disciplinary measures. It is important to recall that our models rely highly on the collected behavior classes that form behavioral sequences over time. Note that the organizations should deeply investigate all behavior classes and ensure what they categorize as misbehaviors before they store them in their database, otherwise false positives may occur, and this may incorrectly indicate the presence of deterioration patterns. In order to avoid false positives and misleading results, organizations should investigate whether people obstruct responsible decision-making and actions. Basically, responsible decision-making refers to the ability to make constructive choices about personal behavior and social interactions based on ethical standards, safety concerns, and social norms. Organizations should also investigate and validate whether reported cases of misbehaviors are genuine or stem from false accusations, accusations of bad faith, and a biased understanding of cultural values, and beliefs. We believe that such investigations will substantially contribute to the precision of the results and the quality of prediction.

Future Work

Several prospective extensions to the problem of behavioral deterioration can be explored, a few of which are detailed below.

- Behavioral deterioration may occur suddenly or slowly, depending upon the pace at which perpetrators cause harm. It is challenging and difficult to detect suddenly deteriorating behaviors. Future work should propose a definition that includes several deterioration scenarios, including the scenario in which the behavioral deterioration occurs suddenly, in order to discover untapped information.

CONCLUSION

- The problem of behavioral deterioration is still in an embryonic phase, it requires significant efforts to be addressed and studied in various facets to reach maturity. In the future, we aim to add multimodal analysis and investigate behavioral sequences without converting misbehavior-related classes into a single class category, and measure our model performance with diverse metrics and datasets. We aim to create the benchmark task for this problem ; a benchmark that will consist of a collection of resources for training, evaluating, and analyzing natural language understanding systems for behavioral deterioration.
- We would like to develop machine learning models to automatically determine the deterioration threshold, that is, a score that could help identify deterioration at a sufficiently early stage to prevent significant further deterioration.
- We would like to build holistic models that combine offline and online inputs. These two inputs could help identify the causes of behavioral deterioration with the help of psychological well-being, social and cultural information, and socioeconomic status.
- We would like to develop pre-trained models that could understand linguistic feature underlying the deterioration task and fine-tune this task based on the nature of data sources.

Publications

List of published works during Ph.D. candidature.

Journal Articles

1. J.M. Tshimula, B. Chikhaoui, and S. Wang. COVID-19 : Detecting depression signals during stay-at-home period. Health Informatics Journal, pages 1-13, 2022.
This work investigates depression signals on Canadian location-specific Twitter data during the first COVID-19 lockdown in Canada, but does not appear as a chapter because we wanted to limit the size of Part I.
2. J.M. Tshimula, B. Chikhaoui, and S. Wang. A new approach for affinity relationships discovery in online forums. Social Network Analysis and Mining Journal, 10, 40, 2020.

Conferences

1. J.M. Tshimula, S. Gray, B. Chikhaoui, and S. Wang. Emotion detection in law enforcement interviews. In Proc. of the 2022 IEEE COMPSAC, HCSC : Human Computing & Social Computing, 2022.
2. J.M. Tshimula, B. Chikhaoui, and S. Wang. Discovering affinity relationships between personality types. In Proc. of the 25th International Conference on Network-Based Information Systems, 2022.
3. J.M. Tshimula, B. Chikhaoui, and S. Wang. Investigating moral foundations from web trending topics. In Proc. of the 25th International Conference on Network-Based Information Systems, 2022.
4. J.M. Tshimula, B. Chikhaoui, and S. Wang. On predicting behavioral deterioration in online discussion forums. In Proc. of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pages 190-195, 2020.
5. J.M. Tshimula, B. Chikhaoui, and S. Wang. A pre-training approach for stance classification in online forums. In Proc. of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pages 280-287, 2020.
6. B. Chikhaoui, J.M. Tshimula, and S. Wang. Community mining and cross-community discovery in online social networks. In Proc. of the 23rd International Conference on Network-Based Information Systems, 2020.
This work does not appear as a chapter because I am not the first author.
7. J.M. Tshimula, B. Chikhaoui, and S. Wang. HAR-search : A method to discover hidden affinity relationships in online communities. In Proc. of 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pages 176-183, 2019.

Bibliography

- [1] R. Abbott, B. Ecker, P. Anand, and M. Walker. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 4445–4452, 2016.
- [2] L. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer. Friendship prediction and homophily in social media. ACM Trans. Web, 6(2), 2012.
- [3] T. Aledavood. Temporal patterns of human behavior. Phd thesis, Aalto University, Department of Computer Science, <https://aaltodoc.aalto.fi/handle/123456789/28892>, 2017.
- [4] M. Alizadeh, I. Weber, C. Cioffi-Revilla, S. Fortunato, and M. Macy. Psychological and personality profiles of political extremists. arXiv preprint arXiv:1704.00119 (Not published paper), 2017.
- [5] T. Allard, R. Wortley, and A. Stewart. The effect of cctv on prisoner misbehavior. The Prison Journal, 88(3):404–422, 2008.
- [6] T. Althoff, D. Borth, J. Hees, and A. Dengel. Analysis and forecasting of trending topics in online media streams. In Proc. of the 21st ACM International Conference on Multimedia, pages 907–916, 2013.
- [7] P. Anand, M. Walker, R. Abbott, J. Tree, R. Bowmani, and M. Minor. Cats rule and dogs drool!: Classifying stance in online debate. In Proc. of the

BIBLIOGRAPHY

- 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pages 1–9, 2011.
- [8] C. Anderson and B. Bushman. Human aggression. Annual Review of Psychology, 53:27–51, 2002.
- [9] M. Antheunis, P. Valkenburg, and J. Peter. The quality of online, offline, and mixed-mode friendships among users of a social networking site. Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 6(3), 2012.
- [10] O. Araque, L. Gatti, and K. Kalimeri. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. Knowledge-Based Systems, 191:105184, 2020.
- [11] E. Aumayr and C. Hayes. On the correlation between topic and user behaviour in online communities. In Proc. of the Tenth International AAAI Conference on Web and Social Media, 2016.
- [12] K. Avrachenkov, N. Litvak, and K. Pham. Distribution of PageRank mass among principle components of the web. In Proc. of the 5th international conference on Algorithms and models for the web-graph, pages 16–28, 2007.
- [13] D. Baras, A. Ronen, and E. Yom-Tov. The effect of social affinity and predictive horizon on churn prediction using diffusion modeling. Social Network Analysis and Mining, 4:232, 2014.
- [14] K. Beune, E. Giebels, W. Adair, and B. Fennis. Strategic sequences in police interviews and the importance of order and cultural fit. Criminal Justice and Behavior, 38(9):934–954, 2011.
- [15] K. Beune, E. Giebels, and K. Sanders. Are you talking to me? influencing behaviour and culture in police interviews. Psychology Crime and Law, 15(7):597–617, 2009.
- [16] J. Biesanz, S. West, and O. Kwok. Personality over time: Methodological approaches to the study of short-term and long-term development and change. Journal of Personality, 71(6):905–41, 2003.

BIBLIOGRAPHY

- [17] P. Binder. Den som vil godt: Om medfølelsens psykologi [the good will: On the psychology of compassion]. Bergen, Germany: Fagbokforlaget, 2014.
- [18] W. Blumenfeld and R. Cooper. Lgbt and allied youth responses to cyberbullying: Policy implications. International Journal of Critical Pedagogy, 3(1):114–133, 2000.
- [19] S. Brânzei and K. Larson. Coalitional affinity games and the stability gap. In Proc. of the Twenty-First International Joint Conference on Artificial Intelligence, pages 79–79, 2009.
- [20] A. Brazinskas, S. Havrylov, and I. Titov. Embedding words as distributions with a Bayesian skip-gram model. In Proc. of the 27th International Conference on Computational Linguistics, pages 1775–1789, 2018.
- [21] I. Briggs Myers and P. Myers. Gifts differing: Understanding personality type. Nicholas Brealy Publishing, 1980.
- [22] J. Briët and P. P. Harremoës. Properties of classical and quantum jensen-shannon divergence. Phys. Rev. A 79, (052311), 2009.
- [23] W. Bukowski and L. Sippola. Friendship and development: Putting the most human relationship in its place. New Dir Child Adolesc Dev., 109:91–98, 2005.
- [24] E. Cambria, S. Poria, R. Bajpai, and B. Schuller. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In Proc. of the 26th International Conference on Computational Linguistics, pages 2666–2677, 2016.
- [25] CAMH. The centre for addiction and mental health (camh). mental health in canada: Covid-19 and beyond. CAMH Policy Advice (July 2020), 2020.
- [26] X. Chai, X. Guo, J. Xiao, and J. Jiang. Analysis of spatial-temporal behavior pattern of the share bike usage during covid-19 pandemic in beijing. Transactions in GIS, 25, number =, 2021.

BIBLIOGRAPHY

- [27] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath. An attentive survey of attention models. ACM Transactions on Intelligent Systems and Technology, 12(5):1–32, 2021.
- [28] M. Chen, J. Liu, and X. Tang. Clustering via random walk hitting time on directed graphs. In Proc. of the Twenty-Third AAAI Conference on Artificial Intelligence, pages 616–621, 2008.
- [29] W. Chen and L. Ku. UTCNN: A deep learning model of stance classification on social media text. In Proc. of the 26th International Conference on Computational Linguistics, pages 1635–1645, 2016.
- [30] Y. Chen, W. Peng, and S. Lee. Mining temporal patterns in time interval-based data. IEEE Transactions on Knowledge and Data Engineering, 27(12):3318–3331, 2015.
- [31] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. Antisocial behavior in online discussion communities. In Proc. of the Ninth ICWSM, pages 61–70, 2015.
- [32] M. Cheong. What are you tweeting about?: A survey of trending topics within the twitter community. Tech. Rep. 2009/251, Clayton School of Information Technology, Monash University (2009), 2009.
- [33] M. Cheong and V. Lee. Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. In Proc. of the 2nd ACM workshop on Social web search and mining, pages 1–8, 2009.
- [34] B. Chikhaoui, M. Chiazzaro, and S. Wang. An improved hybrid recommender system by combining predictions. In Proc. of 2011 IEEE Workshops of International Conference on Advanced Information Networking and Applications, pages 644–649, 2011.
- [35] B. Chikhaoui, M. Chiazzaro, and S. Wang. A new granger causal model for influence evolution in dynamic social networks: The case of dblp. In Proc. of

BIBLIOGRAPHY

- the Twenty-Ninth AAAI Conference on Artificial Intelligence 51, pages 51–57, 2015.
- [36] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734, 2014.
- [37] C. Chung and J. Pennebaker. The psychological functions of function words. Social Communication, pages 343–359, 2007.
- [38] M. Cliche. BBtwtr at SemEval-2017 Task 4: Twitter sentiment analysis with CNNs and LSTMs. In Proc. of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 2017.
- [39] C. Coleman. The Disparate impact argument reconsidered: Making room for justice in the assisted suicide debate. The Journal of Law, Medicine and Ethics, 30(1):17–23, 2002.
- [40] H. Cowie. Cyberbullying and its impact on young people’s emotional health and well-being. The Psychiatrist, 37(5):167–170, 2013.
- [41] T. Davidson, D. Warmley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In Proc. of the Eleventh International AAAI Conference on Web and Social Media, 2017.
- [42] M. Demir and L. Weitekamp. I am so happy ‘cause today i found my friend: Friendship and personality as predictors of happiness. Journal of Happiness Studies, 8:181–211, 2007.
- [43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4171–4186, 2019.

BIBLIOGRAPHY

- [44] L. Dewa, E. Cecil, L. Eastwood, A. Darzi, and P. Aylin. Indicators of deterioration in young adults with serious mental illness: A systematic review protocol. Systematic Reviews, 7:123, 2018.
- [45] K. Dey and S. Majumdar. Customer sentiment analysis by tweet mining: Unigram dependency approach. Indian Journal of Computer Science and Engineering, 6(4):124–129, 2015.
- [46] R. Dhaneriya, M. Ahirwar, and M. Motwani. Unigram polarity estimation of movie reviews using maximum likelihood. International Journal of Computer Science Issues, 13(5):120–124, 2016.
- [47] M. Doroszuk, M. Kupis, and A. Czarna. Personality and friendships. In: Zeigler-Hill V., Shackelford T. (eds) Encyclopedia of Personality and Individual Differences. Springer, Cham., 2019.
- [48] J. Du, R. Xu, Y. He, and L. Gui. Stance classification with target-specific neural attention networks. In Proc. of the Twenty-Sixth International Joint Conference on Artificial Intelligence, pages 3988–3994, 2017.
- [49] J. Ebrahimi, D. Dou, and D. Lowd. A joint sentiment-target-stance model for stance classification in tweets. In Proc. of the 26th International Conference on Computational Linguistic, pages 2656–2665, 2016.
- [50] N. Extremera, C. Quintana-Orts, S. Mérida-López, and L. Rey. Cyberbullying victimization, self-esteem and suicidal ideation in adolescence: Does emotional intelligence play a buffering role? Front Psychol., 9:367, 2018.
- [51] L. Filipović. Police interviews with suspects: Communication problems and possible solutions. Pragmatics and Society, 10(1), 2019.
- [52] F. Fluttert, B. Van Meijel, M. Van Leeuwen, S. Bjørkly, H. Nijman, and M. Grypdonck. The development of the forensic early warning signs of aggression inventory: Preliminary findings: Toward a better management of inpatient aggression. Arch Psychiatr Nurs., 25(2):129–137, 2011.

BIBLIOGRAPHY

- [53] J. Frommel and R. Mandryk. Modeling behaviour to predict user state: Self-reports as ground truth. arXiv preprint arXiv:2007.14461 (Not published paper), 2020.
- [54] A. Furnham, J. Moutafi, and J. Crump. The relationship between the revised neo-personality inventory and the myers-briggs type indicator. Social Behavior and Personality: An International Journal, 31(6):577–584, 2003.
- [55] J. Garten, R. Boghrati, J. Hoover, K. Johnson, and M. Dehghani. Morality between the lines: Detecting moral sentiment in text. In Proc. of IJCAI 2016 Workshop on Computational Modeling of Attitudes, 2016.
- [56] C. Gaskin and G. Dagley. Recognising signs of deterioration in a person’s mental state. Sydney: Australian Commission on Safety and Quality in Health Care, 2018.
- [57] A. Giachanou, P. Rosso, I. Mele, and F. Crestani. Emotional influence prediction of news posts. In Proc. of the Twitter International AAAI Conference on Web and Social Media, pages 592–595, 2018.
- [58] J. Golbeck, C. Robles, M. Edmondson, and K. Turner. Predicting personality from twitter. In Proc. of 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, pages 149–156, 2011.
- [59] L. Gong and H. Wang. When sentiment analysis meets social network: A holistic user behavior modeling in opinionated data. In Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018.
- [60] J. Graham. Morality beyond the lab. Science, 345(6202):1242, 2014.
- [61] A. Granero-Gallegos, R. Baños, A. Baena-Extremera, and M. Martínez-Molina. Analysis of misbehaviors and satisfaction with school in secondary education according to student gender and teaching competence. Frontiers in Psychology, 11(63):1–9, 2020.

BIBLIOGRAPHY

- [62] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks, 18(5-6):602–10, 2005.
- [63] G. Gudjonsson. Psychological vulnerabilities during police interviews. why are they important? Legal and Criminological Psychology, 15:161–175, 2010.
- [64] G. Guibon, M. Ochs, and P. Bellot. From emojis to sentiment analysis. WACAI 2016, Lab-STICC; ENIB; LITIS, Jun 2016, Brest, France. <https://hal-amu.archives-ouvertes.fr/hal-01529708>, 8, 2016.
- [65] S. Guntuku, R. Schneider, A. Pelullo, J. Young, V. Wong, L. Ungar, D. Polsky, K. Volpp, and R. Merchant. Studying expressions of loneliness in individuals using twitter: An observational study. BMJ Open, 9:e030355, 2019.
- [66] S. Guntuku, D. Yaden, M. Kern, L. Ungar, and J. Eichstaedt. Detecting depression and mental illness on social media: An integrative review. Current Opinion in Behavioral Sciences, 18:43–49, 2017.
- [67] T. Guyet and R. Quiniou. Extracting temporal patterns from interval-based sequences. In Proc. of the Twenty-Second International Joint Conference on Artificial Intelligence, pages 1306–1311, 2011.
- [68] J. Haidt and J. Graham. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. Social Justice Research, 20(1):98–115, 2007.
- [69] S. Hajian, T. Tassa, and F. Bonchi. Individual privacy in social influence networks. Social Network Analysis and Mining, 6:2, 2016.
- [70] M. Hamilton. Verbal aggression: Understanding the psychological antecedents and social consequences. Journal of Language and Social Psychology, 31(1):5–12, 2011.
- [71] W. Hamilton, K. Clark, J. Leskovec, and D. Jurafsky. Inducing domain-specific sentiment lexicons from unlabeled corpora. In Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 595–605, 2016.

BIBLIOGRAPHY

- [72] K. Harris and S. Vazire. On friendship development and the big five personality traits. Social and Personality Psychology Compass, 10:647–667, 2016.
- [73] A. Hassan, V. Qazvinian, and D. Radev. What’s with the attitude? Identifying sentences with attitude in online discussions. In Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010.
- [74] K. Haworth. An analysis of police interview discourse and its role(s) in the judicial process. PhD thesis, University of Nottingham, 2009.
- [75] C. Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste. Automatic detection of cyberbullying in social media text. PLOS One, 13(10):e0203794, 2018.
- [76] M. Henderson and A. Furnham. Similarity and attraction: The relationship between personality, beliefs, skills, needs and friendship choice. Journal of Adolescence, 5(2):111–123, 2007.
- [77] A. Herdağdelen, B. State, L. Adamic, and W. Mason. The social ties of immigrant communities in the United States. In Proc. of the 8th ACM Conference on Web Science, pages 78–84, 2016.
- [78] G. Heydon. The language of police interviewing: A critical analysis. Basingstoke: Palgrave, 2005.
- [79] S. Hinduja and J. Patchin. Bullying, cyberbullying, and suicide. Archives of suicide research, 14(3):206–221, 2010.
- [80] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Computation, 9(8):1735–1780, 1997.
- [81] C. Hoffman. Investigative interviewing: Strategies and techniques. International Foundation for Protection Officers. Papers of the Certified Protection Officer Candidates, <https://www.ifpo.org/wp-content/uploads/2013/08/interviewing.pdf>, 2005.

BIBLIOGRAPHY

- [82] U. Holmberg. Crime victims' experiences of police interviews and their inclination to provide or omit information. International Journal of Police Science & Management, 6:155–170, 2004.
- [83] M. Hong, J. Jung, and M. Lee. Social affinity-based group recommender system. ICCASA 2015: Context-Aware Systems and Applications, pages 111–121, 2016.
- [84] I. Hsieh and Y. Chen. Determinants of aggressive behavior: Interactive effects of emotional regulation and inhibitory control. PLoS One, 12(4):e0175651, 2017.
- [85] M. Hu and M. Liu. Mining and summarizing customer reviews. In Proc. of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 168–177, 2004.
- [86] Q. Huang, V. Singh, and P. Atrey. Cyber bullying detection using social and textual analysis. In Proc. of the 3rd International Workshop on Socially-Aware Multimedia, pages 3–6, 2014.
- [87] L. Huesmann and L. Taylor. The role of media violence in violent behavior. Annual Review of Public Health, 27:393–415, 2006.
- [88] Y. Hwang and S. Gelfand. Sparse dynamic time warping. In Proc. of the International Conference on Machine Learning and Data Mining in Pattern Recognition, pages 163–175, 2017.
- [89] A. Hwung. Peer misbehavior effects in the classroom. CMC Senior Theses, http://scholarship.claremont.edu/cmc_theses/1345, page 1345, 2016.
- [90] S. Istia and H. Purnomo. Sentiment analysis of law enforcement performance using support vector machine and k-nearest neighbor. In Proc. of the 2018 3rd International Conference on Information Technology, Information System and Electrical Engineering, 2018.
- [91] D. Jarmolowicz, Y. Hayashi, and C. Pipkin. Temporal patterns of behavior from the scheduling of psychology quizzes. Journal of Applied Behavior Analysis, 43(2):297–301, 2010.

BIBLIOGRAPHY

- [92] N. Jiang and M.-C. de Marneffe. Evaluating BERT for natural language inference: A case study on the commitmentBank. In Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 6086–6091, 2019.
- [93] Q. Jiao, Y. Huang, W. Liu, X. Wang, X. Chen, and H. Shen. Revealing the hidden relationship by sparse modules in complex networks with a large-scale analysis. PLoS ONE, 8(6):e66020, 2013.
- [94] A. John, A. Glendenning, A. Marchant, P. Montgomery, A. Stewart, S. Wood, K. Lloyd, and K. Hawton. Self-harm, suicidal behaviours, and cyberbullying in children and young people: Systematic review. Journal of Medical Internet Research, 20(4):e129, 2018.
- [95] O. John and J. Gross. Healthy and unhealthy emotion regulation: Personality processes, individual differences, and life span development. Journal of Personality, 72(6):1301–1334, 2004.
- [96] O. John and S. Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), Handbook of Personality: Theory and Research, pages 102–138, 1999.
- [97] F. Kachapova. Representing Markov chains with transition diagrams. Journal of Mathematics and Statistics, 9(3):149–154, 2013.
- [98] J. Kahn, R. Tobin, A. Massey, and J. Anderson. Measuring emotional expression with the linguistic inquiry and word count. The American Journal of Psychology, 120(2):263–286, 2007.
- [99] T. Kailath. The divergence and bhattacharyya distance measures in signal selection. IEEE Transactions on Communication Technology, 15(1):52–60, 1967.
- [100] M. Kim and S. Park. Group affinity based social trust model for an intelligent movie recommender system. Journal of Multimedia Tools and Applications, 64(2):505–516, 2011.

BIBLIOGRAPHY

- [101] A. Klein, A. Sarker, M. Rouhizadeh, K. O'Connor, and G. Gonzalez. Detecting personal medication intake in twitter: An annotated corpus and baseline classification system. In Proc. of the BioNLP, Association for Computational Linguistics, pages 136–142, 2017.
- [102] P. Krejzl and J. Steinberger. UWB at SemEval-2016 Task 6: Stance detection. In Proc. of the 10th International Workshop on Semantic Evaluation, pages 408–412, 2016.
- [103] K. Krstovski, D. Smith, H. Wallach, and A. McGregor. Efficient nearest-neighbor search in the probability simplex. In Proc. of the 2013 Conference on the Theory of Information Retrieval, pages 101–108, 2013.
- [104] R. Kumar, A. Ojha, S. Malmasi, and M. Zampieri. Benchmarking aggression identification in social media. In Proc. of the First Workshop on Trolling, Aggression and Cyberbullying, 2018.
- [105] S. Kumar, J. Cheng, and J. Leskovec. Antisocial behavior on the web: Characterization and detection. In Proc. of the 26th International Conference on World Wide Web Companion, 2017.
- [106] I. Kwok and Y. Wang. Locate the hate: Detecting tweets against blacks. In Proc. of the Twenty-Seventh AAAI Conference on Artificial Intelligence, pages 1621–1622, 2013.
- [107] M. Laakasuo, A. Rotkirch, V. Berg, and M. Jokela. The company you keep: Personality and friendship characteristics. Social Psychological and Personality Science, 8(1):66–73, 2016.
- [108] A. Lai, Y. Bisk, and J. Hockenmaier. Natural language inference from multiple premises. In Proc. of the 8th International Joint Conference on Natural Language Processing, pages 100–109, 2017.
- [109] B. Lamiroy and T. Sun. Precision and recall without ground truth. In Proc. of the Ninth IAPR International Workshop on Graphics RECognition - GREC, 2011.

BIBLIOGRAPHY

- [110] J. Lamond, R. Joseph, and D. Proverbs. An exploration of factors affecting the long term psychological impact and deterioration of mental health in flooded households. Environmental Research, 140:325–334, 2015.
- [111] N. Landwehr, M. Hall, and E. Frank. Logistic model trees. In Proc. of the 14th European Conference on Machine Learning, pages 241–252, 2003.
- [112] A. Larsen and M. Crowley. Virtue ethics: Analysing emotions in a police interview with a crime suspect. Policing: A Journal of Policy and Practice, 6(3):291–300, 2012.
- [113] R. Larsen and T. Ketelaar. Personality and susceptibility to positive and negative emotional states. Journal of Personality and Social Psychology, 61(1):132–140, 1991.
- [114] C. Laudy. Hidden relationships discovery through high-level information fusion. In Proc. of 18th International Conference on Information Fusion, pages 916–923, 2015.
- [115] L. Lee. Police interrogation transcripts. Datasets (<https://criminalwords.net/police-interrogation-transcripts/>) posted on September 15, 2019, 2019.
- [116] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, 2020.
- [117] C. Li, A. Porco, and D. Goldwasser. Structured representation learning for online debate stance prediction. In Proc. of the 27th International Conference on Computational Linguistics, pages 3728–3739, 2018.
- [118] Q. Li. New bottle but old wine: A research of cyberbullying in schools. Computers in Human Behavior, 23(4):1777–1791, 2007.

BIBLIOGRAPHY

- [119] T. Li, J. Gharibshah, E. Papalexakis, and M. Faloutsos. Trollspot: Detecting misbehavior in commenting platforms. In Proc. of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pages 171–175, 2017.
- [120] J. Lindert. Cyber-bullying and its impact on mental health. European Journal of Public Health, 27(3), 2017.
- [121] T. Liu, Y. Hu, J. Gao, Y. Sun, and B. Yin. Zero-shot text classification with semantically extended graph convolutional network. In Proc. of the 2020 25th International Conference on Pattern Recognition, pages 8352–8359, 2021.
- [122] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A Robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (Not published paper), 2019.
- [123] M. Luebker. Can the structure of inequality explain fiscal redistribution? Revisiting the social affinity hypothesis. LIS Working papers 762, 2019.
- [124] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder. Hate speech detection: Challenges and solutions. PLoS ONE, 14(8):e0221152, 2019.
- [125] K. Madsen and U. Holmberg. Interviewees’ psychological well-being in investigative interviews: A therapeutic jurisprudential approach. Psychiatry, Psychology and Law, 22(1):60–74, 2014.
- [126] O. Mazni, S. Syed-Abdullah, and N. Hussin. Analyzing personality types to predict team performance. In Proc. of 2010 International Conference on Science and Social Research, pages 624–628, 2010.
- [127] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2):153–157, 1947.
- [128] MHRC. Mental health in crisis: How covid-19 is impacting Canadians. Mental Health Research Canada, 2020.

BIBLIOGRAPHY

- [129] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In Proc. of the 26th International Conference on Neural Information Processing Systems, pages 3111–3119, 2013.
- [130] C. Miller and F. Quek. Interactive data-driven discovery of temporal behavior models from events in media streams. In Proc. of the 20th ACM international conference on Multimedia, pages 459–468, 2012.
- [131] H. Miller, D. Kluver, J. Thebault-Spieker, L. Terveen, and B. Hecht. Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication. In Proc. of the Eleventh International AAAI Conference on Web and Social Media, pages 152–161, 2017.
- [132] R. Morris, M. Carriaga, B. Diamond, N. Piquero, and A. Piquero. Does prison strain lead to prison misbehavior? an application of general strain theory to inmate misconduct. Journal of Criminal Justice, 40(3):194–201, 2012.
- [133] M. Mozafari, R. Farahbakhsh, and N. Crespi. A BERT-based transfer learning approach for hate speech detection in online social media. Computational Intelligence, 881:928–940, 2019.
- [134] E. Mutlu, T. Oghaz, E. Tütüncüler, and I. Garibay. Do bots have moral judgement? the difference between bots and humans in moral rhetoric. In Proc. of 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2020.
- [135] C. Neal. An evaluation of police interviewing methods: A psychological perspective. PhD thesis, Honors Theses, University of Nebraska-Lincoln, 2019.
- [136] S. Needleman and C. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 48(3):443–53, 1970.
- [137] S. Nowson and A. Gill. Look! who’s talking? projection of extraversion across

BIBLIOGRAPHY

- different social contexts. In Proc. of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition, pages 23–26, 2014.
- [138] G. Oh, Y. Zhang, and R. Greenleaf. Measuring geographic sentiment toward police using social media data. American Journal of Criminal Justice, 2021.
- [139] Y. Ozturk. Student misbehavior in the efl classroom: Perceptions of pre- and in-service teachers. Journal of Education and Practice, 8(29):115–122, 2017.
- [140] R. Panigrahy, M. Najork, and Y. Xie. How user behavior is related to social affinity. In Proc. of the 5th ACM international conference on Web search and data mining, pages 713–722, 2012.
- [141] M. Parmar, B. Maturi, J. Dutt, and H. Phate. Sentiment analysis on interview transcripts: An application of nlp for quantitative analysis. In Proc. of the 2018 International Conference on Advances in Computing, Communications and Informatics, 2018.
- [142] J. Pearse, G. Gudjonsson, I. Clare, and S. Rutter. Police interviewing and psychological vulnerabilities: Predicting the likelihood of a confession. Journal of Community & Applied Social Psychology, 8(1):1–21, 1998.
- [143] K. Pearson. Notes on regression and inheritance in the case of two parents. In Proc. of the Royal Society of London, 58:240–242, 1895.
- [144] J. Pennebaker, R. Boyd, K. Jordan, and K. Blackburn. The development and psychometric properties of liwc2015. Austin, TX: University of Texas at Austin, 2015.
- [145] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing, pages 1532–1543, 2014.
- [146] G. Pennycook and D. Rand. Fighting misinformation on social media using crowdsourced judgments of news source quality. In Proc. of the National Academy of Sciences, 116(7):2521–2526, 2019.

BIBLIOGRAPHY

- [147] F. Piedboeuf, P. Langlais, and L. Bourg. Personality extraction through linkedIn. In: Meurs M.J., Rudzicz F. (eds) Advances in Artificial Intelligence. Canadian AI 2019. Lecture Notes in Computer Science, 11489, 2019.
- [148] B. Plank and D. Hovy. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In Proc. of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 92–98, 2015.
- [149] V. Podobnik and I. Lovrek. Implicit social networking: Discovery of hidden relationships, roles and communities among consumers. In Proc. of 19th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Procedia Computer Science 60, pages 583–592, 2015.
- [150] P. Potash and A. Rumshisky. Towards debate automation: A recurrent model for predicting debate winners. In Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2465–2475, 2017.
- [151] I. Price, J. Gifford-Moore, J. Fleming, S. Musker, M. Roichman, G. Sylvain, N. Thain, L. Dixon, and J. Sorensen. Six attributes of unhealthy conversations. In Proc. of the Fourth Workshop on Online Abuse and Harms, pages 114–124, 2020.
- [152] H. Purohit, Y. Ruan, D. Fuhry, S. Parthasarathy, and A. Sheth. On Understanding the divergence of online social group discussion. In Proc. of the Eighth International AAAI Conference on Weblogs and Social Media, pages 396–405, 2014.
- [153] A. Rafaeli and R. Sutton. Expression of emotion as part of the work role. The Academy of Management Review, 12(1):23–37, 1987.
- [154] R. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, and S. Mishra. Scalable and timely detection of cyberbullying in online social networks. In Proc. of the 33rd Annual ACM Symposium on Applied Computing, pages 1738–1747, 2018.

BIBLIOGRAPHY

- [155] A. Razavi, D. Inkpen, S. Uritsky, and S. Matwin. Offensive language detection using multi-level classification. In Proc. of the 23rd Canadian Conference on Advances in Artificial Intelligence, pages 16–27, 2010.
- [156] A. Rezgui, D. Fahey, and I. Smith. AffinityFinder: A system for deriving hidden affinity relationships on twitter utilizing sentiment analysis. In Proc. of the 4th International Conference on Future Internet of Things and Cloud Workshops, pages 212–215, 2016.
- [157] P. Risan, P. Binder, and R. Milne. Emotional intelligence in police interviews—approach, training and the usefulness of the concept. Journal of Forensic Psychology Practice, 16(5):410–424, 2016.
- [158] C. Roberts-Griffin. What is a good friend: A qualitative analysis of desired friendship qualities. Penn McNair Research Journal, 3(1), 2011.
- [159] S. Salawu, Y. He, and J. Lumsden. Approaches to automated detection of cyberbullying: A survey. IEEE Transactions on Affective Computing, pages 1–20, 2017.
- [160] K. Sawhney, M. Prasetio, and S. Paul. Community detection using graph structure and semantic understanding of text. In SNAP Stanford University, 2017.
- [161] M. Selmi. Was the disparate impact theory a mistake? UCLA Law Review, 53(3):701–782, 2006.
- [162] O. Selvitopi, M. Hussain, A. Azad, and A. Buluç. Optimizing high performance markov clustering for pre-exascale architectures. In Proc. of 2020 IEEE International Parallel and Distributed Processing Symposium, pages 116–126, 2020.
- [163] M. Seufert, T. Hoffeld, A. Schwind, V. Burger, and P. Tran-Gia. Group-based communication in whatsapp. In Proc. of 2016 IFIP Networking Conference (IFIP Networking) and Workshops, pages 536–541, 2016.

BIBLIOGRAPHY

- [164] N. Sharma, O. Prakash, K. Sengar, S. Chaudhury, and A. Singh. The relation between emotional intelligence and criminal behavior: A study among convicted criminals. Industrial Psychiatry Journal, 24(1):54–58, 2015.
- [165] A. Shemyakin. Hellinger distance and non-informative priors. Bayesian Analysis, 9(4):923–938, 2014.
- [166] Y. Shih and S. Parthasarathy. Identifying functional modules in interaction networks through overlapping Markov clustering. Bioinformatics, 28(18):473–479, 2012.
- [167] J. Shlens. Notes on kullback-leibler divergence and likelihood theory. arXiv preprint arXiv:1404.2000 (Not published paper), 2014.
- [168] P. Skalny. An application of graph theory in Markov chains reliability analysis. Advances in Electrical and Electronic Engineering, 12(2):154–159, 2014.
- [169] L. Sless, N. Hazon, S. Kraus, and M. Wooldridge. Forming coalitions and facilitating relationships for completing tasks in social networks. In Proc. of the 2014 International Conference on Autonomous Agents and Multi-agent Systems, pages 261–268, 2014.
- [170] L. Smith, L. Zhu, K. Lerman, and Z. Kozareva. The role of social media in the discussion of controversial topics. In Proc. of 2013 International Conference on Social Computing, pages 236–243, 2014.
- [171] T. Smith and M. Waterman. Identification of common molecular subsequences. Journal of Molecular Biology, 147(1):195–197, 1981.
- [172] S. Somasundaran and J. Wiebe. Recognizing stances in ideological on-line debates. In Proc. of the 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pages 116–124, 2010.
- [173] M. Song, W. Lee, and J. Kim. Extraction and visualization of implicit social relations on social networking services. In Proc. of the Twenty-Fourth AAAI Conference on Artificial Intelligence, pages 1425–1430, 2010.

BIBLIOGRAPHY

- [174] D. Sridhar, J. Foulds, B. Huang, L. Getoor, and M. Walker. Joint models of disagreement and stance in online debate. In Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 116–125, 2015.
- [175] D. Sridhar, L. Getoor, and M. Walker. Collective stance classification of posts in online debate forums. In Proc. of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media, pages 109–117, 2014.
- [176] M. Sultana and M. Gavrilova. Temporal pattern in tweeting behavior for persons’ identity verification. In Proc. of 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 2472–2477, 2018.
- [177] B. Sun, Y. Wu, K. Zhao, J. He, L. Yu, H. Yan, and A. Luo. Student class behavior dataset: A video dataset for recognizing, detecting, and captioning students’ behaviors in classroom scenes. Neural Computing and Applications, 33(1):8335–8354, 2021.
- [178] Q. Sun, Z. Wang, Q. Zhu, and G. G. Zhou. Stance detection with hierarchical attention network. In Proc. of the 27th International Conference on Computational Linguistics, pages 2399–2409, 2018.
- [179] R. Sun and D. Shek. Student classroom misbehavior: An exploratory study based on teachers’ perceptions. The Scientific World Journal, page 208907, 2012.
- [180] S. Taheri, H. Mahyar, M. Firouzi, E. Ghalebi, R. Grosu, and A. Movaghar. Extracting implicit social relation for social recommendation techniques in user rating prediction. In Proc. of the 26th International Conference on World Wide Web Companion, pages 1343–1351, 2017.
- [181] L. Tang. Development of online friendship in different social spaces. Information, Communication & Society, 13(4):615–633, 2010.
- [182] A. Thomas, J. Scott, and J. Mellow. The validity of open-source data when assessing jail suicides. Health and Justice, 6(11):1–10, 2018.

BIBLIOGRAPHY

- [183] P. Tieger and B. Barron-Tieger. Just your type. Boston, MA: Little, Brown, and Co., 2000.
- [184] K. Topal, M. Koyutürk, and G. Özsoyoğlu. Effects of emotion and topic area on topic shifts in social media discussions. Social Network Analysis and Mining, 7:46, 2017.
- [185] J. Tshimula, B. Chikhaoui, and S. Wang. HAR-search: A method to discover hidden affinity relationships in online communities. In Proc. of 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pages 176–183, 2019.
- [186] J. Tshimula, B. Chikhaoui, and S. Wang. A new approach for affinity relationship discovery in online forums. Social Network Analysis and Mining, 10(40), 2020.
- [187] J. Tshimula, B. Chikhaoui, and S. Wang. On predicting behavioral deterioration in online discussion forums. In Proc. of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pages 190–195, 2020.
- [188] J. Tshimula, B. Chikhaoui, and S. Wang. Covid-19: Detecting depression signals during stay-at-home period. Health Informatics Journal, pages 1–13, 2022.
- [189] J. Tshimula, B. Chikhaoui, and S. Wang. Discovering affinity relationships between personality types. arXiv preprint arXiv:2202.10437 (Not published paper), 2022.
- [190] M. Tutek, I. Sekulic, P. Gombar, I. Paljak, F. Culinovic, F. Boltuzic, M. Karan, D. Alagic, and J. Snajder. TakeLab at SemEval-2016 Task 6: Stance classification in tweets using a genetic algorithm based ensemble. In Proc. of the 10th International Workshop on Semantic Evaluation, pages 464–468, 2016.
- [191] S. van Dongen. Graph clustering by flow simulation. PhD thesis, University of Utrecht, 2000.

BIBLIOGRAPHY

- [192] Y. Vardi and Y. Wiener. Misbehavior in organizations: A motivational framework. Organization Science, 7(2):151–165, 1996.
- [193] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Proc. of the 31st International Conference on Neural Information Processing Systems, pages 6000–6010, 2017.
- [194] S. Waldman and S. Verga. Countering violent extremism on social media. In Defence Research and Development Canada, 2016.
- [195] M. Walker, P. Anand, R. Abbott, and R. Grant. Stance classification using dialogic properties of persuasion. In Proc. of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 592–596, 2012.
- [196] A. Wang, W. Hamilton, and J. Leskovec. Learning linguistic descriptors of user roles in online communities. In Proc. of the First Workshop on NLP and Computational Social Science, pages 76–85, 2016.
- [197] Z. Waseem, T. Davidson, D. Warmusley, and I. Weber. Understanding abuse: A typology of abusive language detection subtasks. In Proc. of the First Workshop on Abusive Language Online, pages 78–84, 2017.
- [198] T. Wharton. Interjections, language, and the ‘showing/saying’ continuum. Pragmatics & Cognition, 11(1):39–91, 2003.
- [199] A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1112–1122, 2018.
- [200] R. Wilson, K. Harris, and S. Vazire. Personality and friendship satisfaction in daily life: Do everyday social interactions account for individual differences in friendship satisfaction? European Journal of Personality, 29:173–186, 2015.

BIBLIOGRAPHY

- [201] W. Wolny. Emotion analysis of twitter data that use emoticons and emoji ideograms. In Proc. of 25th International Conference on Information Systems Development, pages 476–483, 2016.
- [202] H. Wu. A prediction method of user purchase behavior based on bidirectional long short-term memory neural network model. The 2nd International Conference on Artificial Intelligence and Information Systems, 2021.
- [203] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose. Detecting offensive tweets via topical feature discovery over a large-scale twitter corpus. In Proc. of the 21st ACM international conference on Information and knowledge management, 2012.
- [204] R. Xiang, J. Neville, and M. Rogati. Modeling relationship strength in online social networks. In Proc. of the 19th International Conference on World Wide Web, pages 981–990, 2010.
- [205] W. Yan, H. Zhang, J. Sui, and D. Shen. Deep chronnectome learning via full bidirectional long shortterm memory networks for mci diagnosis. Med Image Comput Comput Assist Interv, 11072:249–257, 2018.
- [206] C. Yang, X. Tang, Q. Dai, and H. Yang. Identifying implicit and explicit relationships through user activities in social media. International Journal of Electronic Commerce, 18(2):73–96, 2013.
- [207] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In Proc. of the fourth ACM international conference on Web search and data mining, pages 177–186, 2011.
- [208] Z. Ye, Y. Geng, J. Chen, J. Chen, X. Xu, S. Zheng, F. Wang, J. Zhang, and H. Chen. Zero-shot text classification via reinforced self-training. In Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3014–3024, 2020.

BIBLIOGRAPHY

- [209] H. Yenala, A. Jhanwar, M. Chinnakotla, and J. Goyal. Deep learning for detecting inappropriate content in text. International Journal of Data Science and Analytics, 6(4):273–286, 2018.
- [210] W. Yin, J. Hay, and D. Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing, pages 3914–3923, 2019.
- [211] W. Yin, J. Hay, and D. Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing, pages 3914–3923, 2019.
- [212] G. Zarrella and A. Marsh. MITRE at SemEval-2016 Task 6: Transfer learning for stance detection. In Proc. of the 10th International Workshop on Semantic Evaluation, pages 458–463, 2016.
- [213] J. Zelenski. The role of personality in emotion, judgment, and decision making. In Vohs, K. D., Baumeister, R. F., & Loewenstein, G. (Eds.), Do emotions help or hurt decision making? A Hedgefoxian perspective, pages 117–132, 2008.
- [214] Z. Zhai, B. Liu, L. Zhang, H. Xu, and P. Jia. Identifying evaluative sentences in online discussions. In Proc. of the Twenty-Fifth AAAI Conference on Artificial Intelligence, 2011.
- [215] A. Zhang, B. Culbertson, and P. Paritosh. Characterizing online discussion using coarse discourse sequences. In Proc. of the Eleventh International AAAI Conference on Web and Social Media, pages 357–366, 2017.
- [216] J. Zhang, J. Chang, C. Danescu-Niculescu-Mizil, L. Dixon, Y. Hua, D. Taraborelli, and N. Thain. Conversations gone awry: Detecting early signs of conversational failure. In Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1350–1361, 2018.
- [217] Z. Zhao, Z. Cheng, L. Hong, and E. Chi. Improving user topic interest profiles by behavior factorization. In Proc. of the 24th International Conference on World Wide Web, pages 1406–1416, 2015.

BIBLIOGRAPHY

- [218] W. Zhou, W. Duan, and S. Piramuthu. A social network matrix for implicit and explicit social network plates. Decision Support Systems, 68(C), 2014.