

SECOND CHANCE COMPETITIVE AUTOENCODERS FOR UNDERSTANDING
TEXTUAL DATA

A Dissertation
IN
Computer Science
and
Electrical Engineering

Presented to the Faculty of the University
of Missouri–Kansas City in partial fulfillment of
the requirements for the degree

DOCTOR OF PHILOSOPHY

by
SARIA GOUDARZVAND

M.Sc. Software Engineering,
Islamic Azad University of Tehran North Tehran Branch, Tehran, Iran, 2014

Kansas City, Missouri
2021

© 2021

SARIA GOUDARZVAND

ALL RIGHTS RESERVED

SECOND CHANCE COMPETITIVE AUTOENCODERS FOR UNDERSTANDING
TEXTUAL DATA

Saria Goudarzvand, Candidate for the Doctor of Philosophy Degree
University of Missouri–Kansas City, 2021

ABSTRACT

Every day, an enormous amount of text data is produced. Sources of text data include news, social media, emails, text messages, medical reports, scientific publications, and fiction. To keep track of this data, there are categories, keywords, tags, or labels that are assigned to each text. Dimensionality reduction and topic modeling in Mining text data has received a lot of attention. Topic modeling is a statistical technique for revealing the underlying semantic structure in a large collection of documents. Applying conventional autoencoders on textual data often results in learning trivial and redundant representations due to high text dimensionality, sparsity, and following power-law word distribution. To address these challenges, we introduce three novel autoencoders, SCAT (Second Chance Autoencoder for Text), SSCAT (Similarity-based SCAT), and CSCAT (Coherent-based SCAT). Our autoencoders utilize competitive learning among the k winner neurons in the bottleneck layer, which become specialized in recognizing specific patterns, leading to learning more semantically meaningful representations of textual data.

In addition, the SSCAT model presents a novel competition based on a similarity measurement to eliminate redundant features. Our experiments prove that SCAT, SSCAT, and CSCAT achieve high performance on several tasks, including classification, topic modeling, compared to LDA, k-Sparse, KATE, NVCTM, ZeroShotTM, and ProDLDA. Additionally, the proposed models are simpler and faster than the established approaches.

This work contributes: (1) SCAT autoencoder utilizes the idea of k-competitive learning among the strongest and weakest, positive, and negative neurons in the bottleneck layer. The novelty stems from involving the weakest neurons in the competition process, which might hold meaningful representations but receive low activation values due to random initialization or being representative of rare words or topics. (2) SSCAT autoencoder presents the novel idea of a similarity-based criterion for selecting neurons that are eligible to enter the learning competition provided by the SCAT approach. This process prevents neurons from a high-similarity score to more than $k/2$ other neurons from entering the competition. We hypothesize that eliminating redundant features will result in better topic representation. (3) CSCAT autoencoder applies the coherent score for selecting the eligible neurons. In this approach we eliminate neurons in which the highest features do not hold a high coherent score. (4) A thorough evaluation of our autoencoders compared to KATE, k-Sparse, LDA and NVCTM. The evaluation includes topic modeling, topic coherence score and document classification using the datasets: 20 Newsgroups, Wiki10+, and Reuters.

APPROVAL PAGE

The faculty listed below, appointed by the Dean of the School of Graduate Studies, have examined a dissertation titled “Second Chance Competitive Autoencoders for Understanding Textual Data,” presented by Saria Goudarzvand, candidate for the Doctor of Philosophy degree, and hereby certify that in their opinion it is worthy of acceptance.

Supervisory Committee

Yugyung Lee, Ph.D., Committee Chair
Department of Computer Science Electrical Engineering
Zhu Li, Ph.D.
Department of Computer Science Electrical Engineering

Deepankar Medhi, Ph.D.
Department of Computer Science Electrical Engineering

Praveen Rao, Ph.D.
Department of Computer Science Electrical Engineering

Reza Derakhshani, Ph.D.
Department of Computer Science Electrical Engineering

Ye Wang, Ph.D.
Department of Communications Studies

CONTENTS

ABSTRACT	iii
ILLUSTRATIONS	ix
TABLES	xii
ACKNOWLEDGEMENTS	xiv
Chapter	
1 INTRODUCTION	1
1.1 Competitive Based Autoencoder For Textual Data	8
1.2 Applications Of Topic Modeling	11
2 SIMILARITY-BASED SECOND CHANCE AUTOENCODER FOR TEXTUAL DATA	17
2.1 Introduction	17
2.2 Related Work	22
2.3 Approach	30
2.4 Experiments	41
2.5 Conclusions	52
3 COHERENCE-BASED SECOND CHANCE AUTOENCODERS FOR DOC- UMENT UNDERSTANDING	58
3.1 Introduction	58
3.2 Related Work	62

3.3	Approach	64
3.4	Experiments	70
3.5	Conclusions	75
4	PERSONALITY TRAIT IDENTIFICATION: CHALLENGES AND OBSTACLES	78
4.1	Introduction	78
4.2	Research Design	80
4.3	Research Methodology	85
4.4	Experimental Results	89
4.5	Discussion	99
4.6	Conclusion	100
5	PUBLIC DISCOURSE ABOUT THE OPIOID CRISIS ON TWITTER, 2010-2019	102
5.1	Introduction	102
5.2	Literature Review	103
5.3	Data And Methods	108
5.4	Results	115
5.5	Discussions	121
5.6	Conclusions	126
6	EARLY TEMPORAL CHARACTERISTICS OF ELDERLY PATIENT COGNITIVE IMPAIRMENT IN ELECTRONIC HEALTH RECORDS	128
6.1	Introduction	128

6.2	Method	131
6.3	Results	135
6.4	Discussion	145
6.5	Conclusion	148
7	MINING NEWS MEDIA FOR UNDERSTANDING PUBLIC HEALTH RECORDS	150
7.1	Introduction	150
7.2	Methods	152
7.3	Results	160
7.4	Discussion	166
7.5	Conclusion	170
8	CONCLUSION	172
8.1	Second Chance Autoencoders For Textual Data	172
8.2	Application Of The Topic Modeling In Different Domain	174
8.3	FUTURE WORKS	177
	Appendix	
	BIBLIOGRAPHY	179
	VITA	211

ILLUSTRATIONS

Figure		Page
1	An example of an autoencoder	3
2	Application of Competitive Based Topic Model	4
3	Example illustrating SSCAT approach. All layers are fully-connected, but the connections are light-colored for illustration purposes.	39
4	Visualization of the 20 Newsgroup documents using T-SNE	46
5	Visualizing the effect of different hyperparameters: K, the number of topics, and the two approaches of similarity measure: Word2Vec versus Topic Vectors.	49
6	Example illustrating the KATE approach [21]. All layers are fully-connected. Values in Red represent activations after the competition.	60
7	Example illustrating CSCAT approach. All layers are fully-connected, but the connections are light-colored for illustration purposes.	65
8	Topic visualization- Religion	73
9	Topic visualization-Politics	73
10	Topic visualization-Sport	73
11	The proposed model (DNN-TFIDF- χ^2)	85
12	Year-by-year Comparison of Death Counts, News Stories, and Tweets . .	111
13	Overall framework of the analysis	113

14	Sentiment of Main Themes of Public Discussion Tweets	117
15	Sentiment of Main Themes of News Story	118
16	Main Themes of public discussion tweets, 2010-2020	119
17	Main Themes of News Story, 2010-2019	120
18	Topics of Personal Experience Tweets	122
19	LIWC Distribution in Personal Experience Tweets	123
20	LIWC Sentiment of Personal Experience Tweets	124
21	Mention of Drugs in Personal Experience Tweets	124
22	Distribution of the first CI diagnosis (CON: consult, SV: subsequent visit, LE: limited exam, ME: multi-system evaluation, SUP: supervisory, SE: specialty evaluation, ADM: admission; GIM: general internal medicine	136
23	Distribution of b-ADL and i-ADL for CI and CU patient groups (x-axis is year(s) before the 1st physicaian-diagnosed CI for CI patients and the latest visit for CU patients; y-axis is a ratio of patients who have a deteriorated ADL)	138
24	ADL distributions for CU and CI patient groups (x-axis is year(s) before the 1st physicaian-diagnosed CI for CI patients and the latest clinical visit for CU patients; y-axis is a ratio of patients who have a deteriorated ADL)	139
25	Topic terms for CI patients - TKM (Experiment 1)	141
26	Topic terms for CI patients - KATE (Experiment 1)	144
27	Topic terms for CI patients - LDA (Experiment 1)	144
28	Topic terms in the TKM model (Experiment 2)	146

29	Topic terms in the KATE model (Experiment 2)	147
30	Topic terms in the LDA model (Experiment 2)	147
31	A schematic view of methods for mining Reuters news. MeSH, Medical Subject Heading; UMLS, Unified Medical Language System	153
32	Normalized numbers of articles and Google Trends searches for the 10 public health issues over time. The numbers are normalized to the highest point on each subfigure. A value of 100 represents the peak popularity for the public health issue.	161
33	Counts of news articles with positive, neutral, and negative sentiments toward 10 public health issues.	163
34	Sentiment scores of news media toward 10 public health issues over 11 years (2007â2017).	164
35	Word clouds of five meaningful topics identified in news articles related to the public health issues, "smoking" and "alcohol drinking."	165

TABLES

Tables	Page
1 Comparative Evaluation with Related Work	54
2 Important Notations	55
3 Training Hyperparameters	55
4 Document classification results of two tasks: Multi-class classification using the 20 Newsgroups dataset and Multi-label classification using both Wiki10+ and Reuters	55
5 Training time in seconds	55
6 <i>p-values</i> for the comparison of classifiers on the 20 Newsgroups dataset, Wiki10+ and Reuters	56
7 Coherence Score Evaluation Results	56
8 Selected Topics in 20 Newsgroup	57
9 datasets	77
10 Important Notations	77
11 Document Classification Results	77
12 Coherence Score Evaluation Results	77
13 Dataset for Experiments	89
14 Overall Experiment Settings	91
15 Facebook Binary Classification for Personality Detection (F1 Score%) . .	94

16	Our proposed model’s Accuracy for Twitter Personality Detection (F1 Score%)	95
17	Facebook Topic-Sentiment Cross-Categorization: DNN-TFIDF- χ^2 Classifier vs. Personality Recognizer	96
18	Twitter Topic-Sentiment Cross-Categorization: DNN-TFIDF- χ^2 Classifier vs. Personality Recognizer	97
19	Opioid News Articles from the New York Times 2010-2019	109
20	Opioid Tweets from 2010-2019	110
21	LIWC Psychological Process	114
22	Topics and Terms	116
23	Average number of clinical notes for CI and CU patients (SD in parenthesis)	135
24	Topic words by TKM (6 months before CI diagnosis)	139
25	Topic words by KATE (6 months before CI diagnosis)	139
26	Topic words by LDA (6 months before CI diagnosis)	140
27	Frequencies of MeSH terms related to public health	154

ACKNOWLEDGEMENTS

Firstly, I wish to express my sincere appreciation to my advisor *Dr.Yugyung Lee*, who convincingly guided and encouraged me to become the best researcher I can and to do the right thing even when the road was tough. Without her persistent mentoring, this dissertation would not have been realized.

I want to thank my committee members for enhancing my research experience. I had the chance to take some of the classes of *Dr.Reza Derakhshani* and *Dr.Zhu Li* and it broaden my knowledge in the area machine learning. I would like to thank *Dr.Praveen Rao* and *Dr. Deep Medhi* for their support and encouragement. Every discussion within planned or just down the hall always provided deeper and interesting perspectives I could incorporate into my research. I started collaborating with *Dr.Ye Wang* in my last years of Ph.D., provided the perfect platform to learn the perfect blend of interdisciplinary research. This collaborating gave me the experience of active academic collaboration.

My summer internship working with professor Sunghwan Sohn at Mayo Clinic was a great experience where I learned a lot about how research gets applied to real-world problems.

I want to thank the School of Computing and Engineering for providing me with an excellent education. Special thanks to my department chair, *Dr. Ghulam Chaudhry* for constant support and guidance.

I want to thank all the UMKC Distributed Intelligent Computing Association (UDICA) members for being present. Over the five years, thank you for creating some

of the best projects I worked on. A few special mentions; *Vijaya Yeruva*, and *Mayanka Chandra Chekar* couldn't have asked better lab mates.

My deep and sincere gratitude to my family for their unparalleled love, help, and support.

I express my gratitude to all the funding agencies; without the financial support completing my dissertation would not be possible. Below is a comprehensive list of funding I have received during of five years of Ph.D.

- UMKC School of Computing and Engineering Teaching Assistant
- UMKC School of Graduate Studies Travel Grants
- UMKC School of Graduate Studies Research Grants
- UMKC Women's Council Graduate Assistance Fund
- Grace Hopper Celebration Scholarship

CHAPTER 1

INTRODUCTION

Topic modeling is a widely used statistical technique for extracting latent variables from huge datasets. It is most suited for text data analysis, although it has also been used to analyze bioinformatics data, social data, health data, and environmental data. This analysis can aid in the organization of large-scale datasets for easier access; for example, structuring databases of journals and articles into groups based on similar focus, social media users into groups based on similar post content, and genomic data into groups based on similar sequence-structure. Topic modeling is a sophisticated text analysis approach that has been used in machine learning, natural language processing (NLP), and data mining for more than two decades. A topic model is used to uncover a set of latent topics in a collection of texts, each of which specifies an interpretable semantic idea. Choosing the right methods for extracting useful statistics and features from a dataset is crucial. Despite the fact that recent topic modeling techniques outperform earlier algorithms, they still require optimization and tweaking in order to produce reliable results [1]. Different topic modeling approaches, as previously indicated, have been created for usage with more specialized data connections and structures, such as short texts [2], long-term sequential data [3], highly correlated data [4], and data with complicated structural links [5]. Recent advances in the field of Machine Learning, especially deep learning [6], [7],

have significantly advanced the state-of-the-art results in several domains, such as computer vision [8], speech recognition [9], and text classification [10]. Unlike traditional Machine Learning algorithms that require a thorough feature engineering phase, neural network-based algorithms can automatically learn the features of the input data and map them to the desired output without human intervention.

Unsupervised learning has been successfully used for detecting patterns from unlabeled data. Understanding large collections of the unstructured text remains a persistent problem. Unsupervised models offer a formalism for exposing a collections themes and have been used to aid information retrieval [11], discover patterns in the medical data sets [12]–[14], video prediction [15]. One of the most popular unsupervised deep learning algorithms is autoencoder [16], [17]. An autoencoder is a neural network [18] that learns data representations by reconstructing the input data at the output layer (i.e., $y^{(i)} = x^{(i)}$). Autoencoders learn the most salient features of the input data by constraining a part of the hidden layers (called the *bottleneck*) often by reducing its dimension less than the input layer. Consequently, the bottleneck neurons become the learned features. Figure 2 illustrates an overview architecture of autoencoders.

Autoencoders have been successfully used for visual applications, such as denoising images [19] and 3D shape retrieval [20]. However, it has been challenging to use autoencoders for textual data due to its high dimensionality and sparsity [21]. Moreover, autoencoders are known to learn trivial representations of textual data due to power-law word distribution [22]. Therefore, to address the challenges mentioned above, several

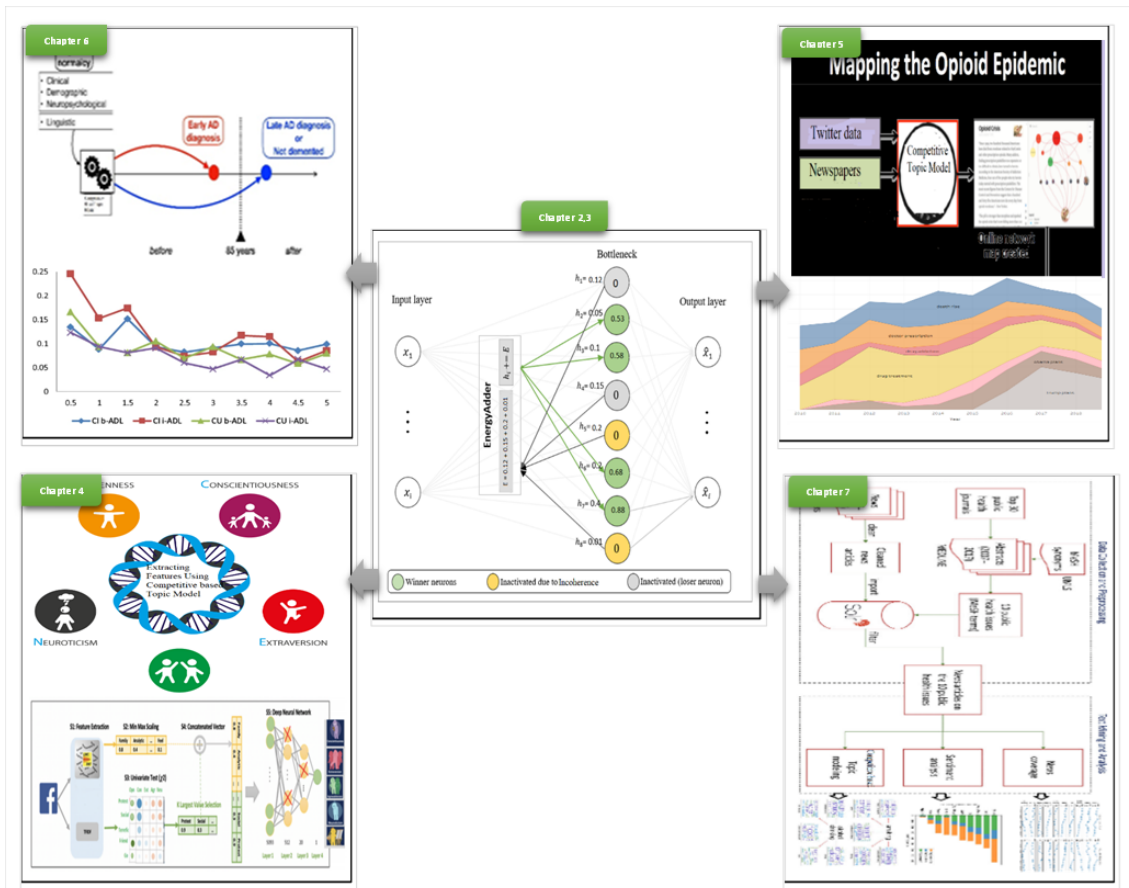


Figure 2: Application of Competitive Based Topic Model

approaches have emerged, including neural autoregressive topic models [23], deep belief networks for topic modeling [24], and neural variational inference for text processing [25], [26].

Other studies introduced the concept of k -competitive autoencoders, which presented outstanding results in the textual data domain, including K-Sparse [27] and KATE [21]. The main idea of k -competitive autoencoders is to select the top k “winner” neurons that then gain the activation values (also referred to as power) from the loser neurons. A competition criterion, such as the largest positive and negative activation values in KATE, is used to select the top k winners. The winner neurons then add the sum of the activation values from the loser neurons to their power while the loser neurons are inactivated, i.e., set to zero.

In k-Sparse, the competition criterion focuses on enforcing sparsity by keeping the k highest activities (activations) during the training and the αk highest activities during validation. However, it is essential to note here that too much sparsity (i.e., a low value of k) can cause the “dead hidden neurons” problem, which results in preventing the weights of these “dead neurons” from being updated in the backpropagation process. The authors of k-Sparse suggested a sparsity scheduling mechanism to mitigate this problem—which introduced extra overhead in the learning process.

KATE (K-competitive Autoencoder for TExt) builds on top of k-Sparse for learning meaningful representations by introducing competition among the neurons of the hidden layers. Specifically, k neurons with the strongest positive and absolute negative activation values gain the power of the rest of the neurons; and thus, they become specialized

in learning more meaningful representations. While KATE does not suffer from the “dead hidden neurons” problem similar to k-sparse, its competition criterion only involves the strongest positive and negative activations of the bottleneck layer (the weakest neurons are loser neurons). In contrast, our study proves that some neurons that hold weak signals in early training cycles can contain important information on useful representative features.

To this end, the three novel autoencoders have been designed to incorporate k-competitive learning: SCAT (Second Chance Autoencoder for Text), SSCAT (Similarity-based SCAT) , and (3) a coherence based (Coherence-based SCAT) filtration method to eliminate neurons that has low coherence (The association between features are small). These autoencoders rely on the idea of k-competitive learning among the neurons of the bottleneck layer so that the k-winner neurons gain the energy of the loser neurons and become specialized in representing meaningful features. Observed from the experiments, it is found that some important words/topics are represented in neurons with small positive and negative activation values. Therefore, they never receive the chance to be represented in the final model. Additionally, it was also observed that conventional k-competitive approaches might include redundant features in the trained model. This challenge has been addressed by constructing the similarity-based autoencoder, SSCAT. Further, another aspect that is significant in topic modeling is that the features per topic has to be coherent. That is, the combinations of top terms per topic should relate to the same concept. This encouraged us to propose CSCAT model. CSCAT filter out neurons that has low coherence score during the training phase.

Consequently, the novelty of our approach stems from proposing a fair k -competitive learning approach that provides (1) a second chance for the *weakest* neurons to reveal their potential, i.e., important topics that otherwise are ignored, (2) a similarity-based filtration technique that eliminates redundant features from the competition process, and (3) a coherence based filtration method to eliminate neurons that has low coherence (The association between features are small). Our approach selects strongest $k/2$ and weakest $k/2$ positive and negative activations as winners. The winner neurons gain the energy of the loser neurons. Note that k is a hyperparameter representing the number of neurons to be included in the competition, and it strongly correlates to the number of topics.

The SSCAT approach presents a novel competition that aims to eliminate redundant features in the learned representations. Specifically, the SSCAT autoencoder prevents neurons with a high-similarity score to more than $k/2$ other neurons from entering the competition. The high similarity is defined by a given threshold, which is equal to the average similarity score across all neurons in the bottleneck layer. The SSCAT design was based on our hypothesis that eliminating redundant features provides a better chance of representing less common features, which is right by our experiments. Similarly, CSCAT has been built on top of the SCAT and filters out neurons that has low coherence score during the training phase.

The primary objectives of the proposed model design were based on the integration of the “second chance” theory and unsupervised learning by utilizing the innovative concept of feeding energy to boost the learning performance of lower performers while

minimizing redundant features during training. It has been shown that the proposed models improved the classification in topic modeling significantly while obtaining comparable results for coherence when compared to the literature.

This dissertation's objectives can be split into two parts: introducing novel competitive based autoencoders for the textual data and the application of topic modeling in the social and health data.

1.1 Competitive Based Autoencoder For Textual Data

The Competitive based autoencoder is done through three different goals.

- SCAT autoencoder utilizes the idea of k -competitive learning among the strongest and weakest, positive, and negative neurons in the bottleneck layer. The novelty stems from involving the weakest neurons in the competition process, which might hold meaningful representations but receive low activation values due to random initialization or being representative of rare words or topics.
- SSCAT autoencoder presents the novel idea of a similarity-based criterion for selecting neurons that are eligible to enter the learning competition provided by the SCAT approach. This process prevents neurons from a high-similarity score to more than $k/2$ other neurons entering the competition. It is hypothesized that eliminating redundant features will result in better topic representation.
- CSCAT autoencoder presents the novel idea of a coherence-based criterion for selecting neurons that are eligible to enter the learning competition provided by the

SCAT approach. This process prevents neurons from a high-similarity score to more than $k/2$ other neurons entering the competition. The higher the association among features per neuron is the more coherent the topic modeling would be.

A brief abstract of each chapter is as follows:

Chapter 2: We introduce Second Chance Autoencoder, a k -competitive learning strategy for textual autoencoders (SCAT). SCAT chooses the k biggest and lowest activations as the winner neurons, which acquire the activation values of the loser neurons during the learning process and hence concentrate on recovering topic-representative information. we introduce a k -competitive autoencoder for extracting meaningful representative features with notable accuracy results. While the k -competitive learning approach was used before in several autoencoders, including K-Spare [27] and KATE [21]; this method differs in the competition criteria. For example, K-Sparse aims at enforcing sparsity in the hidden layers by keeping k highest activities in the training phase and α highest activities in the testing phase (k and α are hyperparameters). The k -competitive approach of KATE selects k winner neurons composed of $k/2$ largest positive activations and $k/2$ largest absolute activations. Then, the k winner neurons gain the energy (i.e., activation value) of the loser neurons. However, through our extensive experiments, we observed that some important words/topics are often represented in neurons with small positive activation values, which are ignored by conventional k -competitive autoencoders. Thus, they never get the chance to be represented in the final model. the novelty of our k -competitive learning approach, SCAT, stems from proposing a fair competition by providing a second chance for the smallest neurons to reveal their potential, i.e., important topics that otherwise are

ignored and hence the name Second Chance Autoencoder (SCAT). Our approach selects the $k/2$ largest (strongest) positive activations and $k/2$ smallest (weakest) positive activations as the k winners, which then gain the energy of the loser neurons. Note that k is a hyperparameter that represents the number of neurons to be included in the competition, and it strongly correlates to the number of topics. Our experiments suggest that setting $k = \frac{\#topics}{2}$ yields higher performance results.

Applying conventional autoencoders for textual data often results in learning trivial and redundant representations due to high text dimensionality, sparsity, and following power-law word distribution. To address these challenges, we propose a novel autoencoders SSCAT (Similarity-based SCAT). Our autoencoder utilize competitive learning among the k winner neurons in the bottleneck layer, which become specialized in recognizing specific patterns, leading to learning more semantically meaningful representations of textual data. Additionally, it was observed that conventional k -competitive approaches might include redundant features in the trained model. This challenge has been addressed by constructing the similarity-based autoencoder, SSCAT. Consequently, the novelty of our approach stems from proposing a fair k -competitive learning approach that provides (1) a second chance for the *weakest* neurons to reveal their potential, i.e., important topics that otherwise are ignored, and (2) a similarity-based filtration technique that eliminates redundant features from the competition process. Our approach selects strongest $k/2$ and weakest $k/2$ positive and negative activations as winners. The winner neurons gain the energy of the loser neurons.

The SSCAT approach presents a novel competition that aims to eliminate redundant features in the learned representations. Specifically, the SSCAT autoencoder prevents neurons with a high-similarity score to more than $k/2$ other neurons from entering the competition. The high similarity is defined by a given threshold, which is equal to the average similarity score across all neurons in the bottleneck layer. The SSCAT design was based on our hypothesis that eliminating redundant features provides a better chance of representing less common features, which is right by our experiments.

Chapter 3:

Often, the terms include in each topics are not highly associated. This will lead to a model that each topic may refer to multiple concept and that makes it difficult for the human to analyze the result. In order to address this challenge, we introduce a new autoencoder, CSCAT (Coherence-based SCAT). Our autoencoder use competitive learning amongst the k winning neurons in the bottleneck layer that become specialized in recognizing specific patterns, leading to learning more semantically significant representations of textual data. In addition, The CSCAT model introduces a new competition based on a measure of consistency to eliminate incoherent features.

1.2 Applications Of Topic Modeling

Topic modeling is a widely used statistical technique for extracting latent variables from huge datasets [28]. It is most suited for text data analysis, although it has also been used to analyze bioinformatics data [29], social data [30], health data [12], and environmental data [31]. This analysis can aid in the organization of large-scale datasets for

easier access; for example, structuring databases of journals and articles into groups based on similar focus [28], social media users into groups based on similar post content [30], and genomic data into groups based on similar sequence-structure [29]. A brief abstract of each chapter related to the application of the topic modeling is as follows:

Chapter 5:

Public discourses on the opioid crisis influence public health policy. Recent studies identify social media as an important public sphere; its agenda and frames reflect ordinary users' perspectives rather than those of health journalists. Thus, the primary purpose of this study is to examine the public discourse about the opioid crisis on Twitter from 2010 to 2019, to inform public health policy-making. The study analyzed 162,760 tweets about the opioid crisis, and compared the main topics and their sentiments with 2,998 opioid stories from nytimes.com. Our analysis separated public discussion tweets from personal experience tweets using clue-based methods. The findings showed that from 2015 to 2019, the top topics of the Twitter discourse were drug addiction, death rate, and epidemic. Twitter's public discourse identified the opioid problem earlier than the news media, and had more diverse and less deliberate perspectives, compared with the news media. The findings suggest that monitoring social media public discourse can raise early warning signs of public concerns over health issues.

Chapter 4: In recent years, information growth has accelerated in lockstep with the emergence of social media, particularly textual data kinds. According to the Social Media Trends study published in [39], there are 3.8 billion active social media users worldwide

as of January 2020, with an annual growth rate of 9.2%. Frequently, individuals use social media to express themselves about a variety of topics, including their personal lives and family well-being, psychology, financial matters, their interactions with societies and the environment, and politics. These terms can be used to define an individual's conduct and personality in various instances. Indeed, prior research (e.g. [32]) has established a substantial association between user personalities and their online social media activity. Personality detection using social media has several challenges. One challenge is that machine learning-based models may not consider the factors that are significant in psychology. Nevertheless, the question is that are manually adding those factors can help in the model building? To answer that, we used Deep Learning techniques and LIWC (Linguistic Inquiry and Word Count) to conduct experimentation. Our goal is to create an optimal model that can predict personality types using social media writing. We conducted experimentation on TFIDF and Chi-square feature selection (TFIDF- χ^2), Doc2Vec, and BERT embedding, with and without LIWC. The results showed that our model achieved higher accuracy than the baseline models without using LIWC. Furthermore, we compared the results achieved by the machine learning-based model with the psychological tool like Personality Recognizer(PR), designed based on stream-of-consciousness writing. We argue that applying machine learning-based models for detecting personality is more suitable for the social media text that is a self-management medium. To compare the result from our model and PR, we studied the relationships among topics, sentiment, and personality types and concluded that the results achieved by machine learning-based models are more in line with theories and evidence from domain fields.

Chapter 6: Another sources of the data that can be used to better analyze the patient's data is the survey patients fill out when they reach out to hospitals. These surveys along with Electronic Health Records (EHRs) can be utilized for the early detection of the diseases. As the population ages, cognitive impairment (CI) has increased, imposing enormous expenses on patients, their families, and society. A large cohort study is urgently needed to better understand the prevalence and severity of CI in order to meet this population's health requirements. However, little is known regarding the temporal patterns in patient health functions (i.e., activities of daily living [ADL]) and the relationship between these trends and the onset of CI in senior patients. Additionally, the utilization of a rich supply of clinical unstructured text inside electronic health records to support CI research has received little attention. The purpose of this study is to describe and better understand early signs of geriatric patient CI by investigating temporal patterns in patient ADL and theme modeling patient medical problems in clinical free text. The study cohort includes of 1,435 physician-diagnosed CI patients ($n = 1,435$) and cognitively unimpaired (CU) patients ($n = 1,435$) who are age and sex matched and were recruited from Mayo Clinic Biobank patients 65 years of age or older at the time of enrolment. A corpus analysis was used to analyze the fundamental statistics of the event types and practice circumstances in which the physician diagnosed CI for the first time. We examined the prevalence of ADL in three distinct age groups prior to the onset of CI. Additionally, we used three distinct topic modeling methodologies using clinical free text to analyze how patients' medical conditions changed over time as they approached CI diagnosis. Based on the experiments' result, 1 to 1.5 year(s) before to the actual physician diagnosis of CI, the

trajectories of ADL deterioration increased steeper in CI patients than in CU patients. The topic modeling revealed that the majority of the topic words were highly connected and accurately described the underlying semantics of CI when approaching CI diagnosis.

Thus, between CI and CU patients, there are significant disparities in the temporal patterns of basic and instrumental ADL. Individual ADL trajectories, such as washing and responsibility for one's own medicine, were found to be strongly related with the development of CI. The subject words extracted from clinical free text using topic modeling techniques have the ability to illustrate how CI patients' problems progress and to disclose previously unrecognized conditions as they approach CI diagnosis.

Chapter 7:

News media play an important role in raising public awareness, framing public opinions, affecting policy formulation, and acknowledgment of public health issues. Traditional qualitative content analysis for news sentiments and focuses are time-consuming and may not efficiently convey sentiments nor the focuses of news media. We used descriptive statistics and state-of-art text mining to conduct sentiment analysis and topic modeling, to efficiently analyze over 3 million Reuters news articles during 2007-2017 for identifying their coverage, sentiments, and focuses for public health issues. Based on the top keywords from public health scientific journals, we identified 10 major public health issues (i.e., air pollution, alcohol drinking, asthma, depression, diet, exercise, obesity, pregnancy, sexual behavior, and smoking). The news coverage for seven public health issues, Smoking, Exercise, Alcohol drinking, Diet, Obesity, Depression, and Asthma decreased over time. The news coverage for Sexual behavior Pregnancy, and Air

pollution fluctuated during 2007â2017. The sentiments of the news articles for three of the public health issues, exercise, alcohol drinking, and diet were predominately positive and associated such as energy. Sentiments for the remaining seven public health issues were mainly negative, linked to negative terms, e.g., diseases. The results of topic modeling reflected the media's focus on public health issues. Thus, text mining methods may address the limitations of traditional qualitative approaches. Using big data to understand public health needs is a novel approach that could help clinical and translational science awards programs focus on community-engaged research efforts to address community priorities.

CHAPTER 2

SIMILARITY-BASED SECOND CHANCE AUTOENCODER FOR TEXTUAL DATA

2.1 Introduction

Recent advances in the field of Machine Learning, especially deep learning [6], [7], have significantly advanced the state-of-the-art results in several domains, such as computer vision [8], speech recognition [9], and text classification [10]. Unlike traditional Machine Learning algorithms that require a thorough feature engineering phase, neural network-based algorithms can automatically learn the features of the input data and map them to the desired output without human intervention.

Unsupervised learning has been successfully used for detecting patterns from unlabeled data. Understanding large collections of the unstructured text remains a persistent problem. Unsupervised models offer a formalism for exposing a collections themes and have been used to aid information retrieval [11], discover patterns in the medical data sets [12]–[14], video prediction [15]. One of the most popular unsupervised deep learning algorithms is autoencoder [16], [17]. An autoencoder is a neural network [18] that learns data representations by reconstructing the input data at the output layer (i.e., $y^{(i)} = x^{(i)}$). Autoencoders learn the most salient features of the input data by constraining a part of the hidden layers (called the *bottleneck*) often by reducing its dimension less than the input layer. Consequently, the bottleneck neurons become the learned features.

Autoencoders have been successfully used for visual applications, such as denoising images [19] and 3D shape retrieval [20]. However, it has been challenging to use autoencoders for textual data due to its high dimensionality and sparsity [21]. Moreover, autoencoders are known to learn trivial representations of textual data due to power-law word distribution [22]. Therefore, to address the challenges mentioned above, several approaches have emerged, including neural autoregressive topic models [23], deep belief networks for topic modeling [24], and neural variational inference for text processing [25], [26].

Other studies introduced the concept of k -competitive autoencoders, which presented outstanding results in the textual data domain, including K-Sparse [27] and KATE [21]. The main idea of k -competitive autoencoders is to select the top k “winner” neurons that then gain the activation values (also referred to as power) from the loser neurons. A competition criterion, such as the largest positive and negative activation values in KATE, is used to select the top k winners. The winner neurons then add the sum of the activation values from the loser neurons to their power while the loser neurons are inactivated, i.e., set to zero.

In k-Sparse, the competition criterion focuses on enforcing sparsity by keeping the k highest activities (activations) during the training and the αk highest activities during validation. However, it is essential to note here that too much sparsity (i.e., a low value of k) can cause the “dead hidden neurons” problem, which results in preventing the weights of these “dead neurons” from being updated in the backpropagation process. The authors of k-Sparse suggested a sparsity scheduling mechanism to mitigate this problem—which

introduced extra overhead in the learning process.

KATE (K-competitive Autoencoder for TExt) builds on top of k-Sparse for learning meaningful representations by introducing competition among the neurons of the hidden layers. Specifically, k neurons with the strongest positive and absolute negative activation values gain the power of the rest of the neurons; and thus, they become specialized in learning more meaningful representations. While KATE does not suffer from the “dead hidden neurons” problem similar to k-sparse, its competition criterion only involves the strongest positive and negative activations of the bottleneck layer (the weakest neurons are loser neurons). In contrast, our study proves that some neurons that hold weak signals in early training cycles can contain important information on useful representative features.

To this end, the two novel autoencoders have been designed to incorporate k-competitive learning: SCAT (Second Chance Autoencoder for Text) and SSCAT (Similarity-based SCAT). Both autoencoders rely on the idea of k-competitive learning among the neurons of the bottleneck layer so that the k-winner neurons gain the energy of the loser neurons and become specialized in representing meaningful features. While the k-competitive learning approach was used before in k-Sparse and KATE, the competition criteria differ from one work to another. For example, KATE’s approach selects k winner neurons composed of $k/2$ largest positive activations and $k/2$ largest absolute negative activations, which then gain the energy of the loser neurons. However, observed from the experiments, it is found that some important words/topics are represented in neurons with small positive and negative activation values. Therefore, they never receive the chance

to be represented in the final model. Additionally, it was also observed that conventional k -competitive approaches might include redundant features in the trained model. This challenge has been addressed by constructing the similarity-based autoencoder, SSCAT.

Consequently, the novelty of our approach stems from proposing a fair k -competitive learning approach that provides (1) a second chance for the *weakest* neurons to reveal their potential, i.e., important topics that otherwise are ignored, and (2) a similarity-based filtration technique that eliminates redundant features from the competition process. Our approach selects strongest $k/2$ and weakest $k/2$ positive and negative activations as winners. The winner neurons gain the energy of the loser neurons. Note that k is a hyperparameter representing the number of neurons to be included in the competition, and it strongly correlates to the number of topics.

The SSCAT approach presents a novel competition that aims to eliminate redundant features in the learned representations. Specifically, the SSCAT autoencoder prevents neurons with a high-similarity score to more than $k/2$ other neurons from entering the competition. The high similarity is defined by a given threshold, which is equal to the average similarity score across all neurons in the bottleneck layer. The SSCAT design was based on our hypothesis that eliminating redundant features provides a better chance of representing less common features, which is right by our experiments.

The proposed models, SCAT and SSCAT, are beyond a simple extension of existing works. The primary objectives of SCAT and SSCAT design were based on the integration of the “second chance” theory and unsupervised learning by utilizing the innovative concept of feeding energy to boost the learning performance of lower performers

while minimizing redundant features during training. It has been shown that the proposed models improved the coherence in topic modeling significantly while obtaining comparable results for classification when compared to the literature. This work contributes:

- SCAT autoencoder utilizes the idea of k -competitive learning among the strongest and weakest, positive, and negative neurons in the bottleneck layer. The novelty stems from involving the weakest neurons in the competition process, which might hold meaningful representations but receive low activation values due to random initialization or being representative of rare words or topics.
- SSCAT autoencoder presents the novel idea of a similarity-based criterion for selecting neurons that are eligible to enter the learning competition provided by the SCAT approach. This process prevents neurons from a high-similarity score to more than $k/2$ other neurons entering the competition. It is hypothesized that eliminating redundant features will result in better topic representation.
- A thorough evaluation of our autoencoders compared to LDA [33], k -Sparse, KATE, ProdLDA [34], NVCTM[35] and ZeroShotTM [36]. The evaluation includes topic coherence score, document classification, and visualization, using the datasets: 20 Newsgroups, Wiki10+, and Reuters.

2.2 Related Work

This section reviews related works in terms of traditional probabilistic topic modeling and deep learning based topic modeling such as Autoencoders and Variational Autoencoders. The summary of the comparative evaluation with the most relevant works is shown in Table 1.

2.2.1 Probabilistic Topic Models

Latent Dirichlet Allocation (LDA) has gained popularity for topic modeling for document collections. The model aims to uncover the latent structure of documents as a mixture of topics by computing a probability distribution over words. Particularly, many variants of LDA have been proposed; non-parametric learning [37], sparsity [38], [39] and efficient inference [40]. The LDA’s main deficiency is that the order of words was not considered due to the underlying usage of “bag of words” [41]. Topic Keyword Model (TKM) was proposed to address this limitation by considering the position word i in a context [42]. TKM fully utilized the critical idea of a joint probability $D \times W$ from the aspect model [43] to highlight certain aspects of the topics in the documents.

In the aspect model, the conditional independence of words w and documents d is assumed given a topic t : $p(d, w) := p(d) \hat{\wedge} p(w|d)$ and $p(w|d) := \sum_t p(w|t) \hat{\wedge} p(t|d)$. TKM conceives the core ideas of the aspect model, but the position i of a word was also considered in text documentations. In addition, the context of a word was taken into account. This implies that each word’s occurrence might have a different probability if a word occurs multiple times in the same document but with other nearby words.

2.2.2 Autoencoders

A basic autoencoder is a shallow neural network that consists of two parts: encoder and decoder. The encoder maps the input layer x to the bottleneck space $z = g(Wx + b)$ while the decoder reconstructs the input at $\hat{x} = g(W^Tz + c)$. Here, W^T is the weight matrix obtained by weight tying, i.e., setting $W\hat{a}^2 = W^T$, which is often used as a regularization method to avoid overfitting, and g , b and c stand for the activation function, the bias term at the encoder, and the bias term at the decoder, respectively.

Zhai et al. [22] discussed that traditional autoencoders perform well on image data but are not good at modeling text documents due to their high dimensionality, sparsity, and following power-law word distribution. Thus, they introduced a semisupervised autoencoder that uses a weight loss function to allow a classifier to learn the network's weights. Their goal was to address the deficiency of the traditional autoencoders when applied to the textual data by introducing a new loss function.

Recent autoencoders that perform well on text classification tasks include k-competitive autoencoders, such as k-Sparse and KATE. The competition criterion is what varies from one method to another. For example, k-Sparse aims to enforce sparsity in the hidden layers by keeping the k highest activities during the training phase and αk highest activities during the validation phase. k-Sparse uses linear activation functions for the hidden neurons, while the non-linearity in the model derives from the selection of k highest activities. k-Sparse achieved better classification results than denoising autoencoders [44], models trained with dropout [45], and Restricted Boltzmann Machines when applied for textual data.

KATE (K-competitive Autoencoder for TExt) builds on top of k-Sparse for learning meaningful representations by introducing competition among the neurons of hidden layers; KATE’s approach is to select k winner neurons composed of $k/2$ largest positive activations and $k/2$ largest absolute negative activations, which then gain the energy of loser neurons. However, due to the competition, a negative neuron may not have had a chance to boost the learning performance of the model if their initial settings are not suitable for learning well or the data smoothing setting is not adequate for optimal learning at the initial stage of training. This means the gap between the negative and the positive is broader and more permanent than expected, forcing an unfair competition situation and resulting in lower overall learning performance.

These works are different from our work in which SCAT and SSCAT encourage sparsity in the model and explore how different sparsity methods will affect the performance of the model. SCAT selects the winner neurons from the strongest and weakest, positive and negative neurons ensuring more fair competition and providing a second chance to the weakest negative and positive neurons. In SSCAT, a filtration technique is introduced that filters out similar neurons before entering the competition process. This ensures that the selected winner neurons are distinctive and not redundant.

2.2.3 Graph Based Topic Model

Graph Neural Networks (GNNs) that capture the interactions between graph nodes via message transmission have been a hot study field in the natural language processing sector. In [46], a GNN based neural topic model that represents a corpus as a document

relationship graph was suggested. Documents and words in the corpus become nodes in the network and are linked based on document-word co-occurrences. By incorporating the graph structure, the links between documents are established by their common words and thus the topical representation of a document is augmented by aggregating information from its surrounding nodes using graph convolution.

In another research [47], they present a novel approach to avoid the overfitting issue of pLSI by employing the amortized inference with word embedding as input, instead of the Dirichlet prior in LDA. For generative topic model, the huge number of free latent variables is the root of overfitting. To decrease the number of parameters, the amortized inference substitutes the inference of latent variable with a function which has the shared (amortized) learnable parameters. The number of the common parameters is constant and irrespective of the magnitude of the corpus. To overcome the restrictive applicability of amortized inference to independent and identically distributed (i.i.d) data, a novel graph neural network, Graph Attention Topic Network (GATON), is developed to represent the topic structure of non-i.i.d documents according to the following two findings. First, pLSI may be viewed as stochastic block model (SBM) on a certain bi-partite network. Second, graph attention network (GAT) may be interpreted as the semi-amortized inference of SBM, which relaxes the i.i.d data assumption of vanilla amortized inference. GATON proposes a unique technique, i.e. graph convolution operation based approach, to merge word similarity and word co-occurrence structure. Specifically, the bag-of-words document representation is treated as a bi-partite graph topology. Furthermore, word embedding, which encodes the word similarity, is modeled as property of the word node and

the term frequency vector is adopted as the attribute of the document node. Based on the weighted (attention) graph convolution operation, the word co-occurrence structure and word similarity patterns are smoothly combined for topic identification

2.2.4 Variational Autoencoders

Generative models have gained notable success for learning from unlabelled data using unsupervised learning. Deep Belief Networks (DBN) is a class of deep generative models that form a direct acyclic graph in which a deep autoencoder is built on the top two layers to reconstruct the input data [48]. DBN could be constructed to reconstruct its input probabilistically. The layers act as a feature detector, and a classifier can be built on top of the detector. Maaloe et al. [24] presented a topic modeling approach based on DBN. The neural variational inference (NVI) approach makes the deep generative framework, such as variational autoencoders, suitable for topic modeling [25]. Neural Variational Document Model (NVDM) is a variational autoencoder-based neural network for document modeling [25]. One downside of NVDM is that it does not consider the correlation among the topics.

Liu et al. [35] proposed Neural Variational Correlated Topic Model (NVCTM), a centralized transformation flow to capture the relationships among topics by reshaping topic distributions. NVCTM consists of the inference network with a centralized transformation flow and a multinomial softmax generative model. The extensive experiments of NVCTM validated its efficiency in capturing perplexity, topic coherence, and document classification tasks. However, although this model often achieves an excellent coherence score, its performance on the classification task is weaker than other related and similar

models, as shown in our results.

The authors in [49] present a topic modeling approach that integrates unsupervised neural topic modeling with supervised neural networks and a novel design in the attention mechanism. Specifically, using a set of documents with labels, this approach uses a variational neural topic modeling model to model a document’s bag of words data and a recurrent neural network classifier to predict the document’s label based on its sequential data. The attention mechanism is used to join the two models. This integration yields a predictive document representation that is suitable for classification. Therefore, while their classification results using the 20news group dataset are higher than ours, our f1 scores on the Wiki dataset are still higher despite their usage of a supervised learning approach. Moreover, our approach can serve more downstream tasks than the classification alone.

An updated version of LDA was introduced in [34] and called ProLDA. This topic model replaces the mixture model used in LDA with a distribution over individual words that are a product of experts. ProLDA produces better topics than standard LDA along with both measurements of topic coherence score and qualitative examination. However, when the model was evaluated based on the accuracy, the performance was not comparable as reported in table 11. The authors also reported efficient usage of the model that could fit a topic model on roughly one million documents in under 80 minutes on a single GPU.

The authors of ProLDA also proposed LDA-VAE, a neural topic model based on the VAE. The normal logistic distribution was employed as the prior over topics for a

topic generation. To further enhance the quality of the generated topic, they replaced the mixture assumption with a weighted product of experts at the word level and proposed the ProdLDA. But both the LDA-VAE and the ProdLDA were not able to produce word-level semantic representations. Besides, the logistic normal prior used in them also could not capture the multiplicity topical aspects in a document and result in generating bad topics.

Recently, a novel neural architecture for cross-lingual topic modeling was proposed; Zero-Shot Topic Model [36]. Zero-Shot Topic Model (ZeroShotTM) is an extension of ProdLDA, and the only difference is that they replaced the BoW with SBERT features [50]. They used a neural encoding layer for the pre-trained document representations from a contextualized embedding model (BERT) before the variational encoder model’s sampling process. The use of language-independent document representations enables a zero-shot topic model on unseen languages. The authors applied their approach to one data set and examined the efficiency of their proposed method on other languages of the same document. This model is an extension of ProdLDA. However, the authors have not discussed how the performance changes if they use a normal variational autoencoder instead of ProdLDA. That is, to examine whether the model’s performance depends on the ProdLDA architecture or introducing SBERT as input to the ProdLDA.

Another approach suitable for large and heavy-tailed vocabularies is the Embedded Topic Model (ETM) [51], a generative model of documents that models each word with a categorical distribution whose natural parameter is the inner product between the word’s embedding and an embedding of its assigned topic. Specifically, topics are modeled as vectors on the word embedding space that uses the dot product between each

word and the topic embedding to define the per-topic distribution over words. However, the ETM cannot analyze a corpus whose topics shift over time.

Wasserstein autoencoders (WAE) and Adversarial-neural Topic Model, named (ATM) are two different topic models based on Wasserstein and Generative Adversarial Network, respectively. WAE [52] regularizes the aggregated posterior to be close to a prior distribution based on the Wasserstein distance. As a result, WAE produces samples of higher quality compared to VAE. The authors of [53] proposed a neural topic model in the Wasserstein autoencoders (WAE) framework, called W-LDA. The W-LDA model enforces Dirichlet prior, which plays a central role in the sparse mixed membership model of LDA. The authors report improvements in topic quality in both coherence and diversity. The authors relate their work to the topic of learning disentangled representations, which can be defined as one where single latent units are sensitive to changes in single generative factors while being relatively invariant to changes in other aspects. This disentanglement illustrates that the topics learned by the model are coherent and distinct—and this argument was supported by [54], which argued that WAE could learn disentangled representation better than VAE.

The authors of [55] created the first Generative Adversarial Network-based Topic Model named ATM. Precisely, the generator network captures the semantic patterns among the latent topics and models the topics using Dirichlet priors. The generator network also produces a representation of word-level semantics. ATM reported improved topical coherence results on datasets NYtimes and Event. However, their overall evaluation is limited to topic coherence. While ATM performs distribution matching in the vocabulary space

with a high dimensional distribution, our approach performs distribution matching in the latent topic space with a smaller space (number of topics), making our approach more efficient to produce higher quality topics than ATM.

Autoencoder based on adversarial training or adversarial autoencoder (AAE) is proposed to replace VAE. The main difference is that while AAE regularizes the aggregated posterior to be close to a prior distribution, the VAE regularizes the posterior to be close to the prior. Some of their recent applications include unaligned text style transfer and semi-supervised natural language inference.

2.3 Approach

Autoencoders draw their technical advantage from constraining part of the hidden layers (i.e., bottleneck), often by reducing its dimensions, to force the neural network to learn representative features from the data and then be used to reconstruct the data at the output layer. However, latent representation layers usually learn the minimal set of trivial, redundant features required to reconstruct the input data. In the case of topic modeling, features are often selected based on the most frequent words following power-law word distributions, which could hinder the overall topic modeling process and thus ignore essential topics associated with less frequent words. Therefore, we propose a competitive learning approach that not only supports the competition among the most significant activation values but also (1) grants a second chance to the neurons with the weakest activations and (2) inactivates the neurons with the highest similarity to other neurons leading to eliminating redundant features. It is shown that introducing sparsity in the model in the form of competitive learning will lead to better performance, especially classification

results.

In our work, the competition criterion arises from a novel finding in the Neuroscience domain, which has already sparked some novel ideas in deep learning. In particular, Mingorance et al. [56] made an exciting finding that the kinase JNK (c-Jun N-terminal protein kinase) provides a *second chance* to the weakest neurons before selecting the neurite that best fits the conditions to form an Axon. Without this fair redistribution of power, weak neurons will never receive a chance to form an Axon. Using this analogy, the k -competitive learning approach was designed to provide the weakest activations (analogous to neurons) a second chance and then select the neurons that activate after energy redistribution (i.e., the second chance). Comprehensive experiments illustrate that such neurons aid the formation of distinctive features. Otherwise, neurons with low activation values—due to random initialization or power-law word distributions—will not be represented in the autoencoder’s latent features without the second chance approach.

Our experiments reflect the findings of [57] from the Neuroscience domain into the deep learning domain and provide evidence towards supporting the correctness of our initial hypothesis—that some essential features might be buried in neurons with low activation values that never receive a chance to appear in the fully-trained network due to initialization randomness or initial low frequency of important words. Based on this idea, these two autoencoders were proposed: SCAT (Second Chance Autoencoders for Text) and SSCAT (Similarity-based SCAT). SSCAT adds another layer of similarity-based competition on top of SCAT. The basic idea of SCAT was presented in our previous work [58]. SSCAT implements the SCAT algorithms and adds on top of it a similarity measurement

to further filter the activations based on their similarities and increase the level of sparsity. In the following, the approach of SSCAT, as mentioned before, encompasses both approaches.

2.3.1 Definitions

SSCAT (Similarity-based SCAT) is defined as a neural network accepting an input vector $x \in R^d$ with d dimensions, and $W \in R^{d \times m}$ is the weight matrix, and h_1, h_2, \dots, h_m are the m hidden neurons, and $\hat{x} \in R^d$ is the output vector. The activation values at the hidden neurons are calculated as $z = g(Wx + b)$, where g represents the activation function and b is the bias at the encoder side. $\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1}$ is defined as the activation function for the hidden neurons and $\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$ as the activation function for the output neurons. The output neurons are defined as $\hat{x} = g(W^T z + c)$, where W^T is the weight matrix obtained by weight tying–sharing–and c is the bias at the decoder side. In this study, the binary cross-entropy loss function, $l(x, \hat{x})$, is used as defined in (3.1), where V is the vocabulary of the dataset.

$$l(x, \hat{x}) = - \sum_{i \in V} x_i \log(\hat{x}_i) + (1 - x_i) \log(1 - \hat{x}_i) \quad (2.1)$$

Given a vocabulary V and the number of times, n_i , a word i is mentioned, the input vectors, x_i , are calculated as given in (3.2).

$$x_i = \frac{\log(1 + n_i)}{\max_{i \in V} \log(1 + n_i)} \text{ for } i \in V \quad (2.2)$$

Algorithm 1: Approach of Training Phase

```
1 procedure Training Phase:
2   for  $e$  in epochs do
3     // feedForward step
4      $z = \tanh(Wx + b)$ 
5     // apply SSCAT
6      $H = \text{sscat\_layer}(k, z)$ 
7     // apply SCAT
8      $H = \text{scat\_layer}(k, H)$ 
9     // apply aggregate
10     $\hat{z} = \text{aggregate\_layer}(k, H)$ 
11    // compute output
12     $\hat{x} = \text{sigmoid}(W^T \hat{z} + c)$ 
13    // compute loss
14     $l = \text{cross\_entropy}(x, \hat{x})$ 
15    // update weights
16    backpropagate_error
17  end
```

Given the model definition, the SSCAT approach goes through the following steps during the training phase at the bottleneck layer (see Algorithm 5): (1) filter out the neurons based on a given similarity measurement, (2) select top k winner neurons, (3) inactivate loser neurons and aggregate their power to the winner neurons, and then continue the regular training process. The SSCAT and SCAT layers' definitions are listed in Algorithms 2 and 3, respectively. Refer to Table 10 a list of notations used in this thesis. Table 3 lists the used hyperparameters for training our models. Each of the steps will be further discussed.

2.3.2 Similarity Filtration Criteria

During the feed-forward step, the SSCAT layer was designed on the hidden neurons z . This novel layer aims to identify neurons that are highly similar to more than $k/2$ other neurons, where the value of k is often set to $m/2$, and m is the number of given topics. It is hypothesized that hidden neurons with high similarity scores are redundant representations or represent similar topics. Thus, similar neurons can be inactivated to provide a better chance of representing less redundant topics.

To identify the most similar neurons, the following two approaches were studied. In the first setting, the corresponding top n most probable words of each neuron were replaced with their equivalent Word2Vec vectors. In the second setting, the similarity among neurons was compared, considering the top n highest activations. In both settings, n represents the number of highest positive values to consider for the similarity measurement. Note that top n activations present a sufficient representation space of the input features distribution; and therefore, restricting the similarity measure to the top n valued units can achieve better results. It was examined running the similarity measure with and without restricting the top n activations. The results encouraged using the top n neurons in both approaches rather than including the entire input features in the similarity measure. It was set as $n = 64$.

After thorough experiments to compare the use of Word2Vec embedding versus raw vector for the similarity measure, we conclude that incorporating Word2Vec achieves better results since it acts like a supervised signal for our algorithm at the beginning of the training and helps the model to inactivate neurons while considering which words per

neuron have a similar meaning. These approaches are elaborated in the following section.

2.3.2.1 Word2Vec Embedding

Let $e \in R^{|V| \times j}$ be the pre-trained word embedding matrix for the vocabulary, rows aligned with W , such that $\|e_i, : \| = 1$, where j is the dimension of the embedding space, and let $e = \{e_1, e_2, \dots, e_n\}$ reflect the embedding vectors of the top n most probable words in the topic space. The median of embedding per topic was computed for the top n word topics E .

Using the topic vectors e , the cosine similarity matrix, $S \in R^{|V| \times m}$, between each vector and the rest of the vectors was calculated as defined in the following equation:

$$S_{ij} = \cos(e_i, e_j) = \frac{e_i \cdot e_j}{\|e_i\| \|e_j\|} \quad (2.3)$$

Where e_i and e_j represent the median embeddings in the topics i and j , respectively. S_{ij} depicts the similarity between the embeddings of topic i and j . Given the similarity score between all the vectors, the mean similarity score was computed across all vectors, which represents the threshold value θ , calculated as follows:

$$\theta = \sum_{i \in m} S_{ij} / m \quad (2.4)$$

Neurons that share a cosine similarity score larger than the threshold with more than $k/2$ neurons are inactivated (i.e., they are disqualified from participating in the learning phase).

Algorithm 2: SSCAT Layer Definition

```
1 function sscat_layer(k, e):  
    // compute cosine similarity for all E (averaged  
    // w2vec correspond to each topic word)  
2 cosine_score = cosine_score(ei, ej)  
    // compute the threshold  
3 θ = mean(cosine_score)  
    // identify highest similar neurons  
    // keep the non-similar neurons  
4 H = count_non_zero(cosine_score  $\leq$  θ)  $\leq$  k/2  
5 return H
```

Therefore, topic vectors that share a high similarity score were filtered using Equation 2.5.

$$H = \text{count_ones}(S_{ij} > \theta) \leq k/2 \text{ for } i, j \in m \quad (2.5)$$

H represents the winner neurons in this layer. Algorithm 2 illustrates the pseudocode of this layer. Note that in the implementation (i.e., source code), H represents a mask matrix that indicates whether or not neurons are inactivated.

2.3.2.2 Vectors Similarity

In the second setting, the vectors of the top n activation values per topic were used for the similarity computation. The value of n was tested against a range of values from 32 to 256, and $n = 64$ yielded the best results for this experiment. Following the same operations conducted in the previous setting, but instead of replacing a word with the Word2Vec vector space, the actual topic vectors were used. It was assumed that V_1, V_2, \dots, V_m , are the vector space of topics, where m is the number of dimensions in

the bottleneck layer, and V_i and V_j are the vector space of the topics i and j respectively, and S_{ij} depicts the similarity between vectors i and j —computed as shown in Equation 2.6. After calculating S_{ij} , θ and H will be computed using Equations 4 and 5 mentioned above.

$$S_{ij} = \cos(V_i, V_j) = \frac{V_i \cdot V_j}{\|V_i\| \|V_j\|} \quad (2.6)$$

Our experiments proved that incorporating supervisory signals, i.e., the word embedding associated with each word of the top n highest activations results in better performance than employing raw vectors from the vector space of the topics since it acts like a supervised signal at the beginning of the training.

2.3.3 Selecting K-Competitive Neurons

After filtering the neurons using their similarity scores obtained in the previous step (using the word2vec embedding similarity measures), the top strongest and weakest, positive and negative activations per dimension m among the eligible vectors were selected. The selected top k neurons, referred to as winners, will gain the activation values of the loser neurons.

In particular, $k/2$ top strongest activation values (positive and negative) and $k/2$ weakest activation values (positive and negative) neurons were selected. Selecting the neurons with the weakest activations in our approach plays a critical role in identifying features that otherwise are buried in weak signals. This is mainly supported by the fact that weak activation values might be caused by (especially in early training epochs): (1) initialization randomness and (2) representing rare (less frequent) words with small values

Algorithm 3: SCAT Layer Definition

```
1 function scat_layer(k, H):  
2    $S_p = \text{get\_strongest\_positive}(k/4, H)$   
3    $S_n = \text{get\_strongest\_negative}(k/4, H)$   
4    $W_p = \text{get\_weakest\_positive}(k/4, H)$   
5    $W_n = \text{get\_weakest\_negative}(k/4, H)$   
6    $H = [S_p, S_n, W_p, W_n]$   
7 return H
```

in the vector space. To ensure that weakest activations have a real potential to become representative features, we track their behavior over training cycles and only keep the neurons that illustrate improvement over time. Otherwise, they are removed from the competition process. For example, lets assume that $z_w^p = \{z_{w_1}, z_{w_2}, \dots, z_{w_k}\}$ is the set of weakest selected neurons in the previous iteration. The values of these activations were re-evaluated, after being considered winners and aggregated new power, to only keep those that grow in power during subsequent iterations, as follows:

$$|z_{w_i}^p| \leq |z_{w_i}^{p+1}| \text{ for } i \in m \quad (2.7)$$

2.3.4 Neuron Power Aggregation

After winner neurons are selected, they add the total activation values from all loser neurons to their current activation value (this process is referred to as neuron power aggregation). Loser neurons are then inactivated (i.e., assigned the value 0). Algorithm 4 defines the final aggregation layer. In the algorithm pseudocode, Z_s and Z_w refer to the sets of positive and negative winner neurons, both strongest and weakest. Note that the

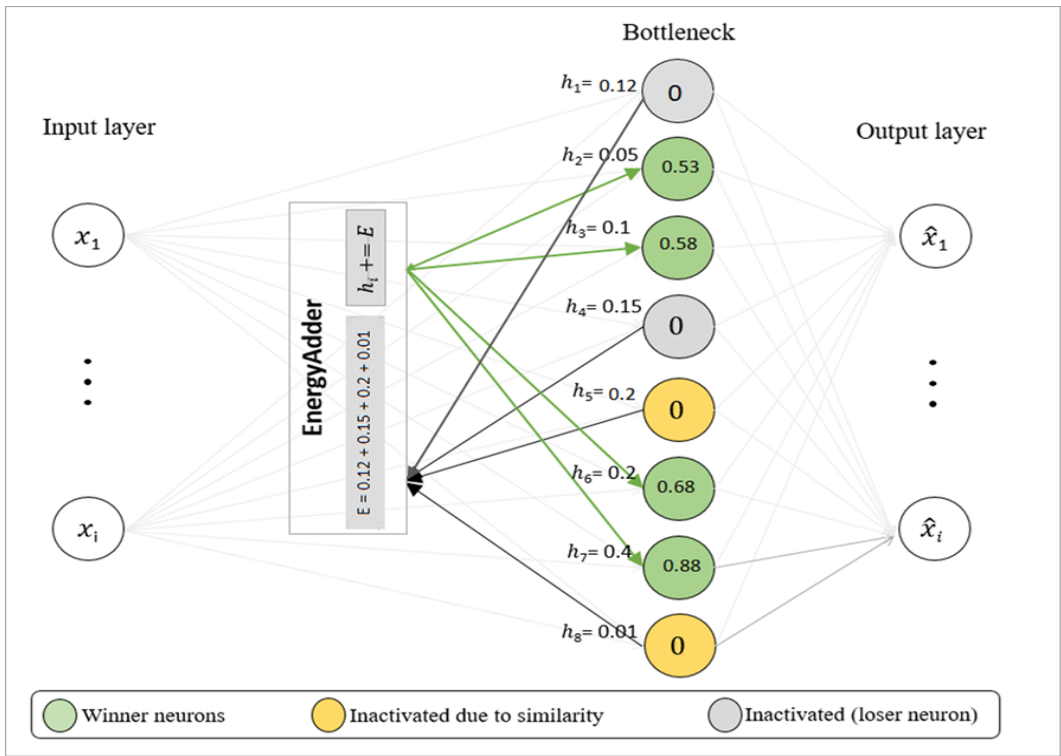


Figure 3: Example illustrating SSCAT approach. All layers are fully-connected, but the connections are light-colored for illustration purposes.

base case scenarios are not included in the algorithms for simplicity.

2.3.5 Illustrative Example

Figure 7 illustrates an example of the training process in SSCAT. For simplicity, the example includes only eight non-negative neurons at the bottleneck layer, and thus it is set as $k = 8/2 = 4$. The original activation values computed by the feedforward step are shown at the top-left of the neurons. After the initialization step, the SSCAT layer identifies and deactivates the neurons with high similarity scores to more than $k/2$ other neurons. It is assumed that neurons h_5 and h_8 have a similarity score higher than the

Algorithm 4: Aggregate Layer Definition

```
1 function aggregate_layer( $k, H$ ):  
    // compute total energy  
2      $E = \sum_{i \in \{z: z = z - [z_s, z_w]\}} z_i$   
    // reallocate energy to winners  
3      $Z_s+ = E$   
4      $Z_w+ = E$   
    // inactivate loser neurons  
5     for  $z$  in  $Z - [Z_s, Z_w]$  do  
6          $z = 0$   
7     end  
8      $\hat{Z} = [Z_s, Z_w, H]$   
9 return  $H$ 
```

threshold with one other neuron, and therefore they are deactivated (colored in Yellow). Then, the rest of the neurons are passed to the SCAT layer, identifying the k winner neurons (colored in Green). Out of the four winners, we select two the highest and two lowest neurons, which are $h_2 = 0.05$, $h_3 = 0.1$ and $h_6 = 0.2$, $h_7 = 0.4$ respectively (no negative activation values were included in this example for simplicity). Then, we sum the energy from the other neurons, including those with high similarity scores (i.e., h_5 and h_8). The total potential energy is $E = 0.12 + 0.15 + 0.2 + 0.01 = 0.48$, and it is added to the winner neurons (i.e., h_2, h_3, h_6, h_7). The rest of the neurons become inactive.

After the feedforward step, we use weight tying to initialize the weights of the hidden-to-output layers. As shown in Algorithm 5, the *sigmoid* activation function was used at the output layer. Then, during the backpropagation step, the gradients will flow through the winner neurons and ignore the loser neurons since they were inactivated. It is important to mention that our algorithm does not require any special steps to encode

data inputs for inference since the network is already well-trained to represent distinctive features and does not require any further processing.

2.4 Experiments

In this section, we evaluate the performance of SCAT and SSCAT on several tasks compared to the current state-of-the-art models. First, the datasets and the relevant baseline models are briefly discussed. All experiments were performed using NVIDIA Titan RTX GPU with 64G RAM. These models were implemented using Keras version 2.2.4 [59] with TensorFlow 1.13 backend [60]. An internal model management tool, called ModelKB, was used to keep track of our experiments [61], [62].

For a fair comparison, we considered the configurations and parameters used in the published results of the baseline algorithms. In terms of the measures, F1, Precision, and Recall were reported based on the results from the evaluation using Monte Carlo cross-validation [63].

2.4.1 Datasets

The three datasets were used in the experiments: 20 Newsgroups [64], Reuters [65], and Wiki10+ [66]. Only the 20 Newsgroups dataset is considered balanced, while the other two are highly imbalanced in terms of class labels. Our results were comparable to the closest work, i.e., KATE [21].

The datasets are randomly partitioned into the training and testing sets using the split (70%/30%) that is typically used in supervised learning studies [67]. However, for

the 20 Newsgroups, the split (60%/40%) was used for the training and testing partition for a fair evaluation with other approaches. We partitioned the 20 Newsgroups, consisting of 18,846 documents over 20 classes, into 60% training set (11,314 documents) and 40% (7,532 documents) test set. The vocabulary included the top 2,000 most frequent words only after removing the stop words and stemming. The Reuters dataset contains 804,414 newswire articles, each with multiple hierarchical topics in 103 categories, into 70% (554,414 articles) training set and 30% (250,000 articles) test set. The vocabulary included the top 5,000 most frequent words. The Wiki10+ dataset contains 19,972 Wikipedia articles with social tags, into about 70% (13,972 articles) training set and 30% (6,000 articles) for the test set. The vocabulary included the top 2,000 most frequent words over 25 tags. In 20 Newsgroups and Wiki10+, 1,000 articles of training data were used for the validation aspect and the Reuters, 10,000 articles for the validation.

2.4.2 Baseline Models

The results of our SCAT and SSCAT models are compared to the following models:

1. LDA [33]: a probabilistic topic model that uses the bag-of-words technique to model a topic and a mixture of topics to model a document.
2. k-Sparse [27]: an autoencoder that enforces sparsity in the hidden layers by keeping the k highest activities in the training phase and the αk highest activities in the validation phase. k-Sparse uses linear activation functions, while the non-linearity in the model derives from the selection of k highest activities.

3. NVCTM [35]: a novel model that proposes the idea of centralized transformation flow to capture the correlations among topics by reshaping topic distributions. Unfortunately, the implementation of this model is not available, so the results reported in their paper were used.
4. ProdLDA [34]: a new topic model that replaces the mixture model in LDA with a product of expert.¹
5. ZeroShotTM [36]: a novel topic model based on ProdLDA in which they used contextualized embedding (e.g., BERT) as input to the model instead of Bow representation. The same setting mentioned in the article was followed while updating the code to support monolingual English topic modeling.²
6. KATE [21]: a shallow autoencoder model with a competitive hidden layer selects the k largest positive neurons and largest absolute negative neurons. Moreover, KATE uses an additional hyperparameter α to amplify the energy value.³

2.4.3 Quantitative Analysis

In this section, we analyze the performance of SCAT and SSCAT compared to the models mentioned above on two tasks: multi-class classification using the dataset of 20 Newsgroups and multi-label classification using the Wiki10+ and Reuters datasets.

¹The ProdLDA source code is available at https://github.com/akashgit/autoencoding_vi_for_topic_models.

²The ZeroShotTM source code is available at <https://github.com/MilaNLProc/contextualized-topic-models>

³The KATE source code is available at <https://github.com/hugochan/KATE>

The results of both tasks are reported in Table 11. We also compare and report the topic coherence scores of these models and support our findings with topic visualizations for the top 10 topics using the 20 Newsgroup dataset compared to KATE, k-Sparse, and LDA.

2.4.3.1 Document Classification

Multi-class classification: This task included training a simple softmax multi-class classifier with a cross-entropy loss function on the 20 Newsgroups dataset. The classification precision, recall, and F1 scores are listed under the 20 Newsgroups column in Table 11. The number of topics is set to 50, and n (number of highest positive activations to consider for the similarity comparison among topic vectors) for all the experiments is set to 64. Changing n from 64 to 500 had little differences in the model’s performance, so we kept $n = 64$ to produce the results in the most efficient time. Note that for the NVCTM model, their published results were used for the comparison, as mentioned previously.

One can observe from the table that competition-based autoencoders achieve better results than conventional models, such as LDA. For example, KATE reaches 70% for all three measurements outperforming NVCTM, k-Sparse, LDA, ProdLDA, and ZeroShotTM, but SCAT and SSCAT outperform all listed models achieving 0.72 scores on all three measures.

Multi-label classification: a multi-label logistic regression classifier was implemented with a cross-entropy loss function to evaluate the models on the multi-label classification task using Wiki10+ and Reuters datasets. The precision, recall, and F1 scores of

these experiments are also listed in Table 11. Note that due to the missing implementation of the NVCTM model, it was not possible to reproduce the results reported in the original paper.

As seen from the table, there are some inconsistencies among the results of this task. We believe that this is because those two datasets, i.e., Wiki10+ and Reuters, are highly-imbalanced. Thus, there exist some differences among the precision on the one hand and the recall on the other hand. KATE wins the precision performance in the Wiki10+ task while SSCAT and SCAT achieve better recall and F1 scores by over 0.3 points. It is also observed that SSCAT and SCAT significantly outperform the rest of the models with the F1 score of 63% for the Reuters dataset.

Statistical Hypothesis Test: In the case of selecting models based on their estimated skill (here F1 score), we are interested to know whether there is a real or statistically significant difference between the baseline model and the rest of the models. We applied the Wilcoxon signed-rank test ⁴, which is a non-parametric statistical hypothesis test used when comparing two paired samples [68], [69]. For simplicity, we compared all the models with SSCAT since SSCAT and SCAT performances did not have statistically significant differences. As we can see in the table 6 in the majority of the cases, SSCAT showed significant differences when compared with other models. SSCAT and ZeroShotTM and SSCAT and KSparse model performances in Wiki10+ and Reuters dataset were less statistically substantial since the p-value is more than 0.05. Also, we have not reported the NVCTM statistical performances since the source code was not available. Note that here

⁴We used the Wilcoxon module in Spacy library: <https://spacy.io/>.

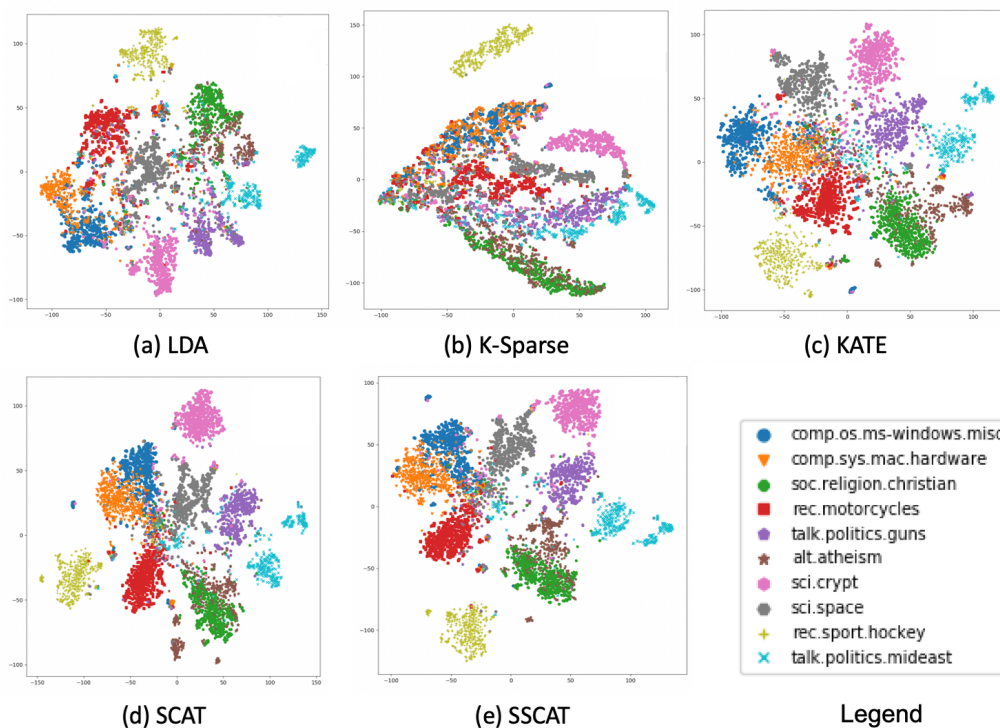


Figure 4: Visualization of the 20 Newsgroup documents using T-SNE

the null hypothesis is that the two models have the same performances. Since, in most cases, the p-value is less than 0.05, the null hypothesis is rejected, and we conclude that the performance of SSCAT is statistically significant.

2.4.3.2 Topic Coherence

We used a topic coherence measurement that is known to have a human-level judgment, called Normalized Point-wise Mutual Information (NPMI) [70]. We evaluated NPMI across all the models using the 20 Newsgroups. We extracted the top-10 words per topic and then computed the NPMI scores as illustrated in Equation 3.6, using topic numbers, $T = 25, 50$ (T is the number of topics).

The results of the NPMI are illustrated in Table 12. We notice that SSCAT achieves scores of 0.141 for 25 topics and 0.118 for 50 topics compared to ProdLDA, with scores of 0.251 and 0.240 for 25 and 50 topics. However, this higher coherence score in ProdLDA comes with a lower classification score compared to both SCAT and SSCAT, as shown in the previous subsection.

$$n\text{pmi}(N) = \sum_{j=2}^N \sum_{i=1}^{j-1} \frac{\frac{\log P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)} \quad (2.8)$$

2.4.3.3 Computational Complexity and Timing

The computational complexity of the proposed SSCAT and SCAT models can be analyzed based on the size and number of trainable parameters present in the autoencoder architectures. The SSCAT and SCAT are autoencoders with competitive learning for feature selection and energy propagation in the negative and positive neurons. The autoencoder network of the SSCAT and SCAT is composed of Encoded Layer (50,025 parameters), Competitive Layer (5,000 parameters), and Decoded Layer (2,000 parameters). Overall, the SSCAT and SCAT have 57,025 parameters. More details of the architecture of the proposed model are described in Table 3.

We also compared the training time of SSCAT and SCAT with the baseline models'. The training times for the 20 newsgroup dataset, with 25 topics, are shown in Table 5. As depicted in the table, the running time is much faster than most baseline models, except LDA and K Sparse.

2.4.4 Qualitative Analysis

In this section, we illustrate that our models can learn semantically meaningful representations from textual data. First, we compare our results to the baseline models, including LDA, k-Sparse, and KATE, using the 20 Newsgroups dataset, with the number of topics set equal to 25. The results are listed in Table 8.

We can observe from the table that our SSCAT model generates the most semantically meaningful topics. Here, we show three topics. The top 10 words learned from the *Religion* category: “resurrection”, “doctrine”, “scripture”, “testament”, “holy”, “jesus”, “spirit”, “christ” and “faith” are strongly related to Religion. SSCAT also learns meaningful representation for the *Sport* category, including words like “players”, “baseball”, “playoffs”, “leafs”, “scoring”, “league”, and “scored” and under *Politics* topic words like “congress”, “senate”, “clinton”, “president”, “secretary”, “administration” which represent the most meaningful representations among the rest of the words generated by other models.

2.4.4.1 Topic Visualization

To further investigate the quality of topic correlation mined by our proposed models, we visualize 10 categories on the 20 Newsgroup dataset compared to KATE, k-Sparse, and LDA. Document visualization is a method to automatically group related documents and visualize their clusters. Particularly, we used the t-SNE (t-distributed Stochastic Neighbor Embedding) [71] visualization, which is well-suited for visualizing high-dimensional data in a low-dimensional space in which similar objects are visualized

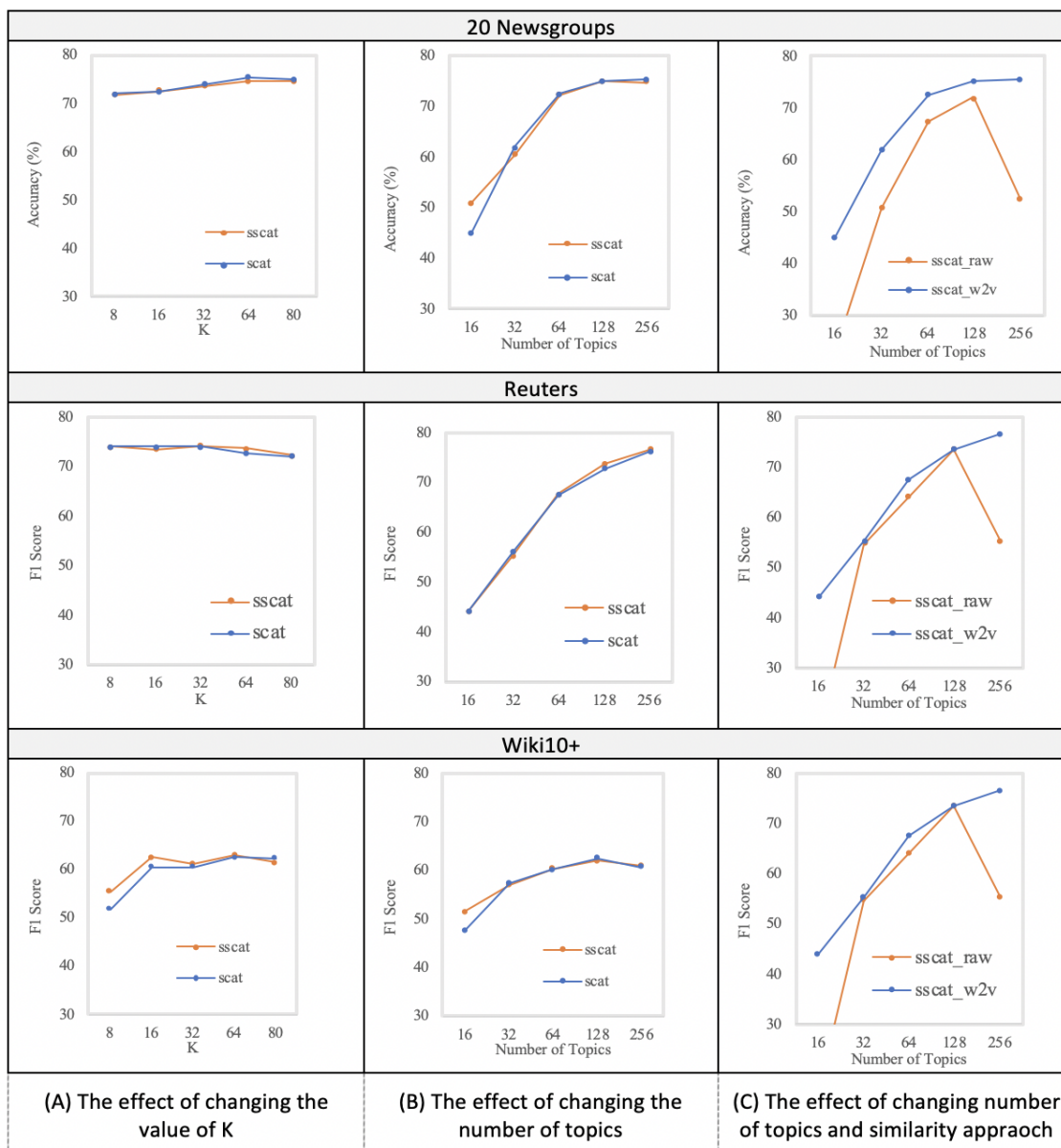


Figure 5: Visualizing the effect of different hyperparameters: K, the number of topics, and the two approaches of similarity measure: Word2Vec versus Topic Vectors.

by nearby points. In contrast, dissimilar objects are visualized using distant points. We set $k = 12$, $\alpha = 6.26$ in KATE and $k = 25$ in SCAT and SSCAT.

Figure 4 depicts the visualization results of five models: LDA, k-Sparse, KATE, SCAT, and SSCAT. The number of topics for all the models is set to 50, and only 10 categories are visualized. The points with different shapes and colors indicate various topics. It can be observed that the proposed models achieve better data representation results for grouping similar topics together. Furthermore, one can easily distinguish between the groups and categories, as shown in Figure 4. This validates the effectiveness of the proposed models in learning meaningful representation from textual data.

2.4.5 Discussions

The experiments verify that our proposed models achieve high-performance results close to or better than the current state-of-the-art results. Specifically, the SCAT and SSCAT models outperform the baseline models in the multi-class classification, multi-label classification, and document visualization tasks. In addition, while the topic coherence score is still best achieved by the ProDLDA model, the SSCAT coherence score achieves a comparable result. Moreover, the qualitative analysis illustrates that our models learn better topic representations than the other models.

Moreover, compared to other approaches that require setting multiple additional parameters to the learning process, such as α and k in KATE, our approach requires setting a single parameter only, i.e., k , which reduces the overhead of parameter search task. Also, ZeroShotTM requires pre-trained representation to feed to the model. Additionally,

hlit can be facilitated that task by recommending setting the value of k to half the number of the given topics (i.e., the number of neurons in the bottleneck layer). To validate this recommendation, it was experimenting with the models' performance over different topic values, as shown in Figure 5(A). With the number of topics set to 128 in all datasets, setting $k = 64$ yielded the best results in both autoencoders. Notably, it was observed that increasing the number of topics did not generally improve the overall performance of the models.

It was observed from Figure 5(B) that the performance of both models gradually increases as the number of topics increases from 20 to 256 while keeping the rest of the hyperparameter unchanged—except for the value of k . However, it was noted that the performance does not significantly change when increasing the number of topics beyond 256, and in some cases, it actually starts to decrease. We argue here that the needed computational powers to create valuable models using several topics beyond 256 requires expensive training experiments but results in a performance improvement compared to the results obtained when training on half the number of topics as shown in Figure 5(B).

One of the significant differences between our approach and KATE is that the latter selects the largest positive and largest absolute negative neurons as winner neurons. In contrast, our approach selects the largest and smallest positive and negative neurons. This critical difference stems from our observations that some meaningful representations can be buried in neurons with low values, either due to random initialization or becoming representative for features with rare words. Thus such neurons never receive a chance to represent the corresponding features. Such neurons become overpowered by neurons

with strong signals—even those with negative values.

Another significant difference from previous similar approaches is that our filtration technique first selects which neurons qualify to enter the competition before choosing the k winner neurons in the actual competition. As discussed in Subsection 3.2, our approach uses a similarity criterion between the word2vec embeddings of the training data. This allows the elimination of most redundant features and provides a space for less frequent features to be represented in the final latent features. We studied two different approaches to achieve this: word2vec embeddings and the similarity between the actual vectors of the input text. Figure 5(C) illustrates the difference in our model’s performance using both similarity approaches. It is noted from the results that using the similarity measure on the embedding vectors achieves better performance results on the underlying tasks using all three datasets. Thus, this technique helped eliminate redundant features allowing less frequent ones to be represented in latent representations.

2.5 Conclusions

We proposed two novel autoencoders for textual data, named SCAT and SSCAT. The models are built upon the idea of k -competitive learning, in which a k winner neurons participate in the learning process while the rest of the neurons are inactivated. Through competition, the winning neurons become highly specialized in learning distinctive features. Unlike previous approaches that introduced the competition between the strongest positive and negative neurons, our approach first eliminates the highly similar neurons. It then presents a completion for the highest and lowest, positive and negative neurons in

the bottleneck layer of the autoencoder.

Our experiments validated that our approach achieves very close or better performance results on several textual data applications, including classification, topic modeling, and document visualization. Moreover, our models provide more semantically essential topics than the baseline models. Because of this, the proposed model is suitable for dimensionality reduction for textual data.

Table 1: Comparative Evaluation with Related Work

Work	Approach	Limitations
LDA [33]	generative statistical model; Each topic is modeled as an infinite mixture over an underlying set of topic probabilities.	If the true structure is more complex than a multinomial distribution or if the data to train isn't sufficient, then it might underfit.
k-Sparse [27]	An autoencoder with linear activation function, where in hidden layers only the k highest activities are kept	Because of the dead neuron issue they used Scheduling of the Sparsity Level which may make it difficult to set the parameters when applying on different dataset.
ProdLDA [34]	LDA-VAE, a neural topic model based on the VAE	Unable to support word-level semantic representations & the multiplicity topical aspects
KATE [21]	Extension of k-Sparse with competitive learning among the neurons of hidden layers	Potential cases of unfair competition between negative and positive neurons. Also, the model needs to set different parameters including alpha.
ZeroShotTM [36]	Extension of ProdLDA with SBERT embeddings for zero shot learning	The model has been built on top of ProdLDA which has already achieved good result. They have not shown what the comparison of their model with ProdLDA on different metrics.
SCAT (ours)	A competition based autoencoder which gives the second chance to the weakest neurons.	It does not consider and examine the quality of the learned features in the training phase.
SSCAT (ours)	A competition based autoencoder that not only gives the second chance to the weakest neurons but eliminates the neurons with high similarity from participating in the training phase.	We have been trying to overcome the limitations of the existing works. In addition to the similarity score, we may need to examine the effectiveness of incorporating other methods, including the co-occurrence of correlation scores.

Table 2: Important Notations

Notation	Description
m	Number of Dimension (Topics)
k	Number of winner neurons
j	Dimension of pre-trained $w2vec$
z_s	Set of strongest activations
z_w	Set of weakest activations
e	Embedding per topic
n	Number of highest activations per topic
H	List of distinctive topics
E	Energy of the activations

Table 3: Training Hyperparameters

Hyperparameter	Value
Number of epochs	50
n	64
j	128
optimizer	adam
Multi-class activation	softmax
Multi-label activation	sigmoid

Table 4: Document classification results of two tasks: Multi-class classification using the 20 Newsgroups dataset and Multi-label classification using both Wiki10+ and Reuters

Model	20 Newsgroups			Wiki10+			Reuters		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
LDA [33]	0.42	0.50	0.46	0.72	0.45	0.56	0.70	0.49	0.44
k-Sparse [27]	0.42	0.42	0.42	0.72	0.45	0.56	0.80	0.50	0.62
NVCTM [35]	0.57	0.56	0.57	-	-	-	-	-	-
ProdLDA [34]	0.53	0.53	0.53	0.49	0.45	0.47	0.51	0.46	0.49
ZeroShotTM [36]	0.63	0.63	0.63	0.58	0.55	0.59	0.55	0.56	0.57
KATE [21]	0.70	0.70	0.70	0.73	0.45	0.56	0.70	0.49	0.61
SCAT (ours)	0.72	0.72	0.72	0.66	0.53	0.59	0.81	0.52	0.63
SSCAT (ours)	0.72	0.72	0.72	0.63	0.58	0.60	0.80	0.52	0.63

Table 5: Training time in seconds

Model	LDA	K-sparse	NVCTM	ProdLDA	ZeroShotTM	KATE	SSCAT	SCAT
Training time (sec)	401	495	-	915	2792	665	550	525

Table 6: p -values for the comparison of classifiers on the 20 Newsgroups dataset, Wiki10+ and Reuters

20 Newsgroups			Wiki10+			Reuters		
Model1	Model2	p -value	Model1	Model2	p -value	Model1	Model2	p -value
SSCAT	SCAT	1.55e-6	SSCAT	SCAT	4.32e-3	SSCAT	SCAT	1.32e-1
SSCAT	LDA	1.65e-6	SSCAT	LDA	8.43e-5	SSCAT	LDA	8.36e-5
SSCAT	KSparse	1.65e-6	SSCAT	KSparse	7.95e-5	SSCAT	KSparse	5.36e-1
SSCAT	ProdLDA	1.64e-6	SSCAT	ProdLDA	8.33e-5	SSCAT	ProdLDA	8.30e-5
SSCAT	ZeroShotTM	1.66e-6	SSCAT	ZeroShotTM	1.50e-1	SSCAT	ZeroShotTM	2.81e-4
SSCAT	KATE	4.52e-3	SSCAT	KATE	1.21e-4	SSCAT	KATE	4.00e-2

Table 7: Coherence Score Evaluation Results

Model	T = 25	T = 50
LDA [33]	0.112	0.140
k-Sparse [27]	0.093	0.090
NVCTM [35]	0.180	0.176
ProdLDA [34]	0.251	0.240
ZeroShotTM [36]	0.159	0.147
KATE [21]	0.073	0.101
SCAT (ours)	0.113	0.104
SSCAT (ours)	0.141	0.118

Table 8: Selected Topics in 20 Newsgroup

Category	Model	Topics
Religion	LDA [33]	god people jesus Christian subject bible church Christ time life
	k-Sparse [27]	god world people origin subject pad Christian bottom application mind
	ProdLDA [34]	holy moral god spirit time christian sale mac faith christ
	ZeroShotTM [36]	people michael spiritual subject jesus life car drive pious christ
	KATE [21]	god michael rutgers dod jesus christian bike drive uga christ
	SCAT (ours)	resurrection doctrine scripture christianity tes- tament christians jesus christ bible faith
	SSCAT (ours)	resurrection doctrine scripture testament holy jesus spirit christ faith christians
Politics	LDA [33]	armenian turkish armenians people war turkey muslim muslims armenia turks
	k-Sparse [27]	israel ca card system israeli national govern- ment state american lawfound
	ProdLDA [34]	armenian genocide turks turkish muslim mas- sacre turkey armenians armenia greek
	ZeroShotTM [36]	armenian writes civilians turks death govern- ment military israeli policy issue
	KATE [21]	article writes nsa gov news org israel israeli
	SCAT (ours)	hitler civilians villages militiary deaths hum- burg miller martin lebanese zone
	SSCAT (ours)	congress senate clinton president batf compound secretary administration waco stephanopoulos
Sport	LDA [33]	Game team year subject games hockey players play writes good
	k-Sparse [27]	team steve good mike players season hockey in- ternet win article
	ProdLDA [34]	season nhl team hockey playoff puck league fly- ers defensive player
	ZeroShotTM [36]	team games sale good subject play scoring dis- count monthly hockey
	KATE [21]	game games hockey team red win season nhl play leafs
	SCAT (ours)	leafs playoffs stanley scoring hockey season team scored clipper
	SSCAT (ours)	baseball players playoffs leafs scoring scored league espn lemieux islanders

CHAPTER 3

COHERENCE-BASED SECOND CHANCE AUTOENCODERS FOR DOCUMENT UNDERSTANDING

3.1 Introduction

Deep neural networks [18] have revolutionized many domains, especially unstructured data, including computer vision [8], speech recognition [9], and text classification [10] to name a few. While most current neural network applications use supervised learning, unsupervised learning has also presented significant advances in extracting patterns in unlabeled data with reasonable efficiency. For example, unsupervised models have been used to aid information retrieval [11], discover patterns in medical datasets [12], [13], and video prediction [15].

One of the most popular unsupervised deep learning algorithms are autoencoders [16], [17]. An autoencoder is a neural network that learns data representations by reconstructing the input data at the output layer (i.e., $y^{(i)} = x^{(i)}$), where $y^{(i)}$ is the network's output (prediction) for the $x^{(i)}$ input sample. Thus, the main objective for autoencoders is to learn the important features of the input data by constraining the size of the middle layer named *bottleneck*, often by reducing its dimension less than the input layer.

While autoencoders have demonstrated significant results in several domains, most notably visual applications such as image compression **huszar2020lossy** and denoising images [19]; it has been challenging to use autoencoders for textual data due to the text

high-dimensionality and sparsity [21]. Adding to the aforementioned challenges, autoencoders are also known to learn trivial representations of the text due to the power-law word distribution [22].

Fortunately, the research community has identified the need for proper methods to utilize autoencoders for textual-data applications, leading to the emergence of several methods that address these challenges. This new body of research introduced several innovations, including neural autoregressive topic models [23], deep belief networks for topic modeling [24], and neural variational inference for text processing [25], [26]. Additionally, in order to better understand textual data and learn more semantically meaningful representations, other research established the idea of k -competitive autoencoders, which produced impressive results in the textual data domain, including *K-sparse* [27] and *KATE* (K-competitive Autoencoder for TExt) [21] depicted in Figure 6.

The primary principle behind k -competitive autoencoders is to select the top k "winner" neurons that conquer the activation values (a.k.a. *power*) from the loser neurons. By incorporating competition among the neurons of the hidden layers, these methods aim to specialize the winner neurons in learning meaningful representations of the text. The top k winners are chosen based on some competition criteria.

For instance, *K-sparse* focuses on maintaining sparsity by preserving the k highest activations during training and the αk highest activations during testing, where α is a hyperparameter. Similarly, *KATE* selects k winners made up of the $\lceil k/2 \rceil$ largest positive activations and the $\lfloor k/2 \rfloor$ largest absolute negative activations. Those winner neurons then acquire the total energy of the loser neurons, which become inactive, i.e., set to zero.

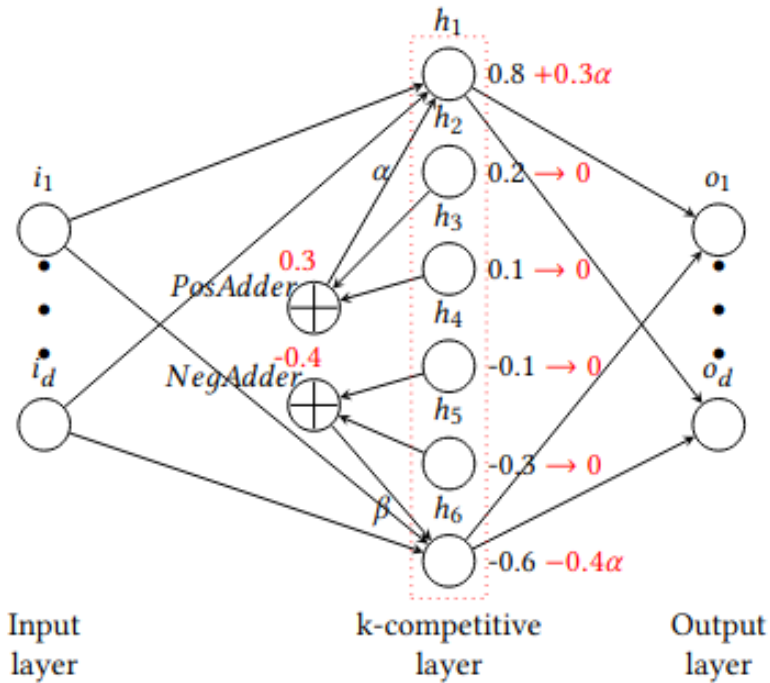


Figure 6: Example illustrating the KATE approach [21]. All layers are fully-connected. Values in Red represent activations after the competition.

k is a hyperparameter representing the desired number of neurons to compete, and it is strongly related to the number of topics.

K-sparse is vulnerable to the "dead hidden neurons" problem caused by adding too much sparsity (low k values), and therefore some neurons can never be updated in the back-propagation process. While this issue can be addressed by incorporating a sparsity scheduling technique, this solution adds significant overhead during the learning process. In contrast, KATE was built on top of K-sparse and solved the dead hidden neurons problem. However, its competition considers the largest positive and negative activations (the weakest neurons are loser neurons) only—leading to ignoring some essential knowledge preserved in low signal neurons that are never selected as winners. Indeed, our research

proves that some of the neurons that maintain weak signals during early training cycles might hold important information on representative features.

To this end, we present CSCAT (Coherence-based SCAT), a novel autoencoder that builds on earlier work in k-competitive learning called SCAT [58]. CSCAT achieves two main innovations over the previous k-competitive learning methods. First, it provides a second chance for the *weakest* neurons to reveal their potential, i.e., important topics that would otherwise be ignored. Second, a coherence-based filtration technique that removes non-coherent neurons from the competition process. Our extensive evaluation demonstrates that these two innovations can lead to better results compared to the prior work in this domain. To summarize, our work contributes the following:

- A novel idea of a coherence-based criterion for filtering neurons that are eligible to enter the learning competition produced by the SCAT layer. This process prevents neurons from a low-coherence score to more than $k/2$ other neurons entering the competition. We hypothesize that eliminating not coherent features during the training phase will result in better topic representations.
- A thorough evaluation and comparison to KATE, K-sparse, LDA [33], NVCTM [35] and ProdLDA. The evaluation tasks include topic modeling, topic coherence score, document classification, and visualization using three datasets: 20 News-groups, Wiki10+, and Reuters dataset.

3.2 Related Work

For topic modeling of document collections, Latent Dirichlet Allocation (LDA) has gained prominence. By constructing a probability distribution across words, the model seeks to reveal the hidden structure of documents as a combination of topics. Non-parametric learning [37], sparsity [38], [39] and efficient inference [40] are only a few of the LDA versions that have been developed. The fundamental flaw in the LDA is that the order of words was not taken into account because of the underlying use of "bag of words" [41]. To solve this issue, the Topic Keyword Model (TKM) was created, which takes into account the position word i in a context [42]. TKM fully utilized the critical idea of a joint probability $D \times W$ from the aspect model [43] to highlight certain aspects of the topics in the documents. TKM conceives the main ideas of the aspect model, but in text documentations, the position i of a word was also taken into consideration. A word's context was taken into account. This means that if a word appears repeatedly in the same document but with different neighboring words, each occurrence may have a different probability. In [34], a new version of LDA called ProLDA was released. This topic model substitutes the mixture model used in LDA with a product of expert distribution across particular words. In terms of topic coherence score and qualitative assessment, ProLDA creates better topics than regular LDA. When the model was tested based on accuracy, however, the results were not similar, as shown in table 11.

Even with ideal reconstructions, autoencoders often only extract simple representations of text data; however, by adding proper regularization to the models, more meaningful representations can be generated. Many autoencoder versions have lately been proposed based on this premise [27], [72]. K-competitive autoencoders, such as KATE, are recent autoencoders that perform well on text classification tasks. KATE (K-competitive Autoencoder for TExt) builds on K-Sparse for learning meaningful representations by introducing competition among hidden layer neurons. KATE’s approach is to select k winner neurons composed of $\text{ceil}k/2$ largest positive activations and $\text{floor}k/2$ largest absolute negative activations, which then gain the energy of loser neurons.

Overall, CSCAT’s technique is fairly similar to that of traditional k-competitive autoencoders. However, we choose the winners from among the strongest and weakest positive and negative neurons, guaranteeing more equal competition and giving the weakest negative and positive neurons a second chance. Second, before starting the competitive process, we offer a filtration mechanism that filters out incoherent neurons. This guarantees that the winning neurons are distinctive and coherent.

Unsupervised learning has seen a lot of success with generative models for learning from unlabeled data. Deep Belief Networks (DBN) are a type of deep generative model in which the input data is reconstructed using a deep autoencoder based on the top two layers of a directed acyclic graph [48]. Maaloe et al. [24] introduced a topic modeling approach based on DBN. The neural variational inference (NVI) approach makes the deep generative framework, such as variational autoencoders, suitable for topic modeling [25].

Neural Variational Document Model (NVDM) is a variational autoencoder based neural network for document modeling [25]. One disadvantage of NVDM is that it ignores the correlation between the topics. Liu et al. [35] presented the Neural Variational Correlated Topic Model (NVCTM), a centralized transformation mechanism that reshapes topic distributions to express links between topics. NVCTM consists of two components: the inference network with a centralized transformation flow and a multinomial softmax generative model. NVCTM’s efficiency in capturing perplexity, topic coherence, and document categorization tasks has been proven through rigorous testing. Although this model frequently earns a high coherence score, its classification performance is inferior to that of other related and similar models.

3.3 Approach

Autoencoders draw their technical advantage from constraining bottleneck, often by reducing its dimensions, to force the neural network to learn representative features from the data, and then be used to reconstruct the data at the output layer. However, latent representation layers usually learn the minimal set of trivial, redundant features required to reconstruct the input data. When it comes to topic modeling, features are frequently chosen based on the most common words based on power-law word distributions, which might block the whole process and lead to the omission of important topics linked with less frequent terms. As a result, we propose a competitive learning approach that not only encourages the competition among the most significant activation values but also (1) gives a second chance to the neurons with the weakest activations and (2) inactivates the

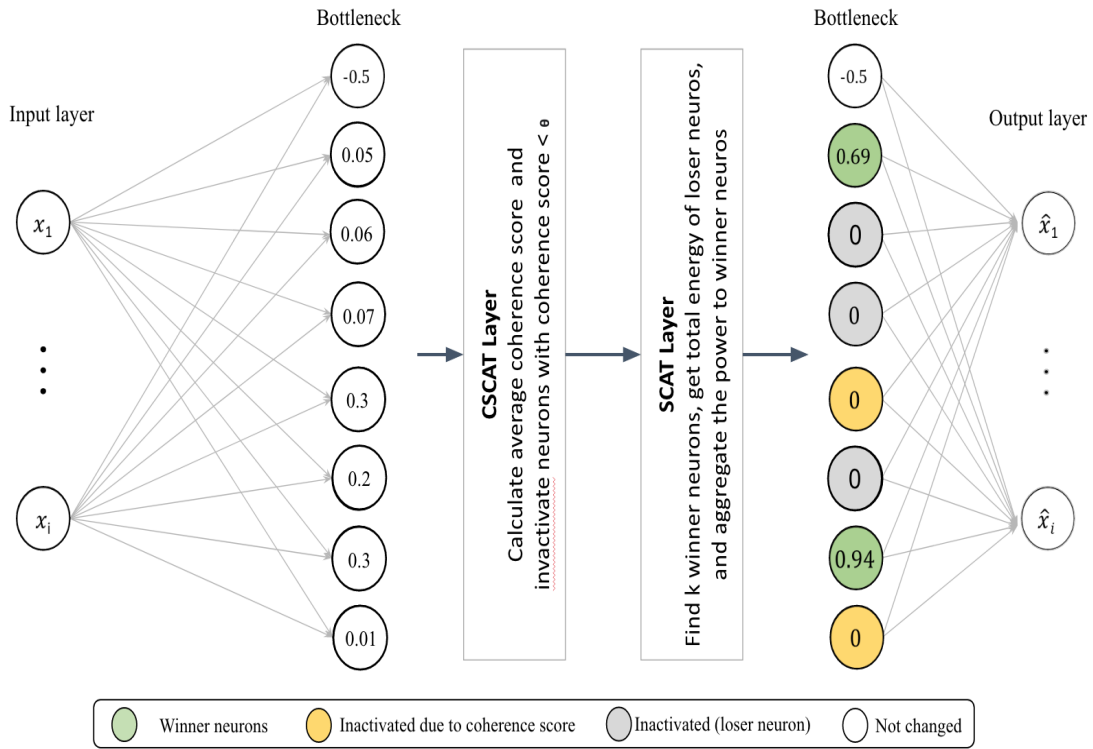


Figure 7: Example illustrating CSCAT approach. All layers are fully-connected, but the connections are light-colored for illustration purposes.

neurons with the lowest coherence during training phase. Figure 7 illustrates an example of the training process in CSCAT.

The competition criterion in our study is based on a unique finding in the Neuroscience area that has already spawned numerous novel deep learning approaches. Mingo-rance et al. [56] discovered that the kinase JNK (c-Jun N-terminal protein kinase) gives the weaker neurons a *second chance* before choosing the neurite that best meets the criteria to produce an Axon. Weak neurons will never have a chance to form an Axon unless there is a fair allocation of power. Without this fair redistribution of power, weak neurons

will never receive a chance to form an Axon. Using this analogy, we designed our k -competitive learning approach to provide the weakest activations (analogous to neurons) a second chance and then selecting the neurons that activate after energy redistribution (i.e., the second chance). Without the second chance technique, neurons with low activation values due to random initialization or power-law word distributions will not be reflected in the autoencoder’s latent features.

Our experiments reflect the findings of [57] from the Neuroscience domain into the deep learning domain and prove the correctness of our initial hypothesis—that some essential features might be buried in neurons with low activation values that never receive a chance to appear in the fully-trained network due to initialization randomness or initial low frequency of important words. Based on this idea, we present a autoencoder: CSCAT (coherence-based SCAT). CSCAT adds another layer of coherence-based competition on top of SCAT (Second Chance Autoencoder). In the following, we explain the approach of CSCAT.

3.3.1 Definitions

We define CSCAT as a neural network accepting an input vector $x \in \mathbb{R}^d$ with d -dimensions, and $W \in \mathbb{R}^{d \times m}$ is the weight matrix, and h_1, h_2, \dots, h_m are the m hidden neurons, and $\hat{x} \in \mathbb{R}^d$ is the output vector. The activation values at the hidden neurons are calculated as $z = g(Wx + b)$, where g represents the activation function and b is the bias at the encoder side. We use $\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1}$ as the activation function for the hidden neurons and $\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$ as the activation function for the output neurons. The

output neurons are defined as $\hat{x} = g(W^T z + c)$, where W^T is the weight matrix obtained by weight tying–sharing–and c is the bias at the decoder side. We use the binary cross-entropy loss function, $l(x, \hat{x})$, as defined in (3.1), where V is the vocabulary of the dataset.

$$l(x, \hat{x}) = - \sum_{i \in V} x_i \log(\hat{x}_i) + (1 - x_i) \log(1 - \hat{x}_i) \quad (3.1)$$

Given a vocabulary V and the number of times, n_i , a word i is mentioned, the input vectors, x_i , are calculated as given in (3.2).

$$x_i = \frac{\log(1 + n_i)}{\max_{i \in V} \log(1 + n_i)} \text{ for } i \in V \quad (3.2)$$

Given our model definition, the CSCAT approach goes through the following steps during the training phase at the bottleneck layer (see Algorithm 5): (1) filter out the neurons based on a given coherence measurement, (2) select top k winner neurons, (3) inactivate loser neurons and aggregate their power to the winner neurons, and then continue the regular training process. Refer to Table 10 a list of notations used. We further explain each of the steps, as mentioned earlier in the following.

3.3.2 Coherence-Based Rule

One of the major issues in clustering particularly in topic modeling is that the final topic words are not coherent. In fact, the association among the top words per topic could be a good indication of the highly correlated words. In training phase, We want to ensure that the words learned by the model are logically consistent per topic. Point-wise mutual

information **bouma2009normalized** is one measure of the statistical independence of observing two words in close proximity. Given a learned W , the practice to extract top-N most probable words for each topic is to take the most positive entries in each column. We define the topic coherence metric NPMI **aletras2013evaluating** as

$$npmi(T_i) = \sum_{j=2}^n \sum_{i=1}^{j-1} \frac{\frac{\log P(T_{w_i}, T_{w_j})}{P(T_{w_i})P(T_{w_j})}}{-\log P(T_{w_i}, T_{w_j})} \quad (3.3)$$

where T_{w_i} and T_{w_j} are the topic word i and j in the sets of filtered topics. w is the list of top-N words for a topic. For a model generating m topics, the overall npmi score is an average over all topics. However, since we incorporate the coherence score into the training phase, we consider the coherence of each topic separately. Thus, top-N word of each topic T_i will get an score (coherence). This score will be compared with the *mean* of scores, and the topics that have coherence score less than the mean will get inactivated during training phase. This process helps in eliminating topics that may not lead to a coherent topic.

$$average_c = \sum_{i=1}^m \frac{npmi(T_i)}{m} \quad (3.4)$$

where m is number of topics or dimension. Thus, we compare each $npmi(T_i)$ with $average_c$ and select those that meet our condition; having coherence greater or equal the average of coherence across all topics.

3.3.3 Selecting K-Competitive Neurons

After filtering the neurons using their coherence scores obtained in the previous step, we select the top strongest and weakest, positive and negative activations per dimension m among the eligible vectors. The selected top k neurons, referred to as winners, will gain the activation values of the loser neurons.

In particular, we select $k/2$ top strongest activation values (positive and negative) and $k/2$ weakest activation values (positive and negative) neurons. Selecting the neurons with the weakest activations in our approach plays a critical role in identifying features that otherwise are buried in weak signals. This is mainly supported by the fact that weak activation values might be caused by (especially in early training epochs): (1) initialization randomness and (2) representing rare (less frequent) words with small values in the vector space. To ensure that weakest activations have a real potential to become representative features, we track their behavior over training cycles and only keep the neurons that illustrate improvement over time. Otherwise, they are removed from the competition process. For example, let's assume that $z_w^p = \{z_{w_1}, z_{w_2}, \dots, z_{w_k}\}$ is the set of weakest selected neurons in the previous iteration. We will re-evaluate the values of these activations, after being considered winners and aggregated new power, to only keep those that grow in power during subsequent iterations, as follows:

$$|z_{w_i}^p| \leq |z_{w_i}^{p+1}| \text{ for } i \in m \quad (3.5)$$

3.3.4 Neuron Power Aggregation

After winner neurons are selected, they add the total activation values from all loser neurons to their current activation value (we refer to this process as neuron power aggregation). Loser neurons are then inactivated (i.e., assigned the value 0). Algorithm 4 defines the final aggregation layer. In the algorithm pseudocode, Z_s and Z_w refer to the sets of positive and negative winner neurons, both strongest and weakest. Note that the base case scenarios are not included in the algorithms for simplicity.

3.4 Experiments

In this section, we evaluate the performance of CSCAT on several tasks compared to the current state-of-the-art models. First, we briefly discuss the used datasets and the relevant baseline models. All experiments were performed using Nvidia Titan RTX GPU with 64G RAM. We implemented our models using Keras version 2.2.4 [59] with TensorFlow 1.13 backend [60]. We used an internal model management tool, called ModelKB, to keep track of our experiments [61], [62]. We used three datasets in our experiments: 20 Newsgroups [64], Reuters [65], and Wiki10+ [66]. The details about the datasets can be found in table 9

3.4.1 Baseline Models

The results of our CSCAT model is compared to the following models:

1. LDA [33]: a probabilistic topic model that uses the bag-of-words technique to model a topic and a mixture of topics to model a document.

2. K-Sparse [27]: an autoencoder that enforces sparsity in the hidden layers by keeping the k highest activities in the training phase and the αk highest activities in the testing phase. K-Sparse uses linear activation functions, while the non-linearity in the model derives from the selection of k highest activities.
3. NVCTM [35]: a novel model that proposes the idea of centralized transformation flow to capture the correlations among topics by reshaping topic distributions. The implementation of this model is not available, so we compared our results to the results reported in their original paper.
4. KATE [21]: a shallow autoencoder model with a competitive hidden layer selects the k largest positive neurons and largest absolute negative neurons. Moreover, KATE uses an additional hyperparameter α to amplify the energy value.
5. ProdLDA [34]: a new topic model that replaces the mixture model in LDA with a product of expert.

3.4.2 Quantitative Analysis

In this section, we analyze the performance of CSCAT compared to the models mentioned above on two tasks: multi-class classification using the dataset of 20 News-groups and multi-label classification using the Wiki10+ and Reuters datasets. The results of both tasks are reported in Table 11. We also compare and report the topic coherence scores of these models.

3.4.2.1 Multi-class classification

This task included training a simple softmax multi-class classifier with a cross-entropy loss function on the 20 Newsgroups dataset. The classification precision, recall, and F1 scores are listed under the 20 Newsgroups column in Table 11. We set the number of topics to 50, and n (number of highest positive activations to consider for the coherence comparison among topic vectors) for all the experiments is set to 15. Changing n from 10 to 50 had little differences in the model’s performance, so we kept $n = 15$ that we achieved best result. Also, note that we did not run the experiment on the NVCTM model, rather we obtained these results from its research paper, as discussed above.

It is obvious from the table that competition-based autoencoders achieve better results than conventional models, such as LDA. For example, KATE achieves 70% for all three measurements outperforming NVCTM, k-Sparse, and LDA, but CSCAT outperform all listed models achieving 0.72 scores on all three measures.

3.4.2.2 Multi-label classification:

we implemented a multi-label logistic regression classifier with a cross-entropy loss function to evaluate the models on the multi-label classification task using Wiki10+ and Reuters datasets. The precision, recall, and F1 scores of these experiments are also listed in Table 11. Note that due to the missing implementation of the NVCTM model, we could not reproduce the results reported in the original paper.

We observe from the table that there are some inconsistencies among the results of this task. We believe that this is because those two datasets, i.e., Wiki10+ and Reuters,

are highly-imbalanced. Thus, there exist some differences among the precision on the one hand and the recall on the other hand. KATE wins the precision accuracy in the Wiki10+ task while CSCAT wins both the recall and F1 scores. We also observe that CSCAT significantly outperform the rest of the models.



Figure 8: Topic visualization- Religion



Figure 9: Topic visualization- Politics



Figure 10: Topic visualization- Sport

3.4.2.3 Topic Coherence

We used a topic coherence measurement that is known to have a human-level judgment, called Normalized Point-wise Mutual Information (NPMI) [70]. We evaluated NPMI across all the models using the 20 Newsgroups. We extracted the top-10 words

per topic and then computed the NPMI scores as illustrated in Equation (3.6), using topic numbers, $T = 25, 50$.

The results of the NPMI are illustrated in Table 12. We notice that CSCAT achieves scores of 0.151 for 25 topics and 0.118 for 50 topics compared to NVCTM, which scores of 0.180 and 0.176 for 25 and 50 topics, consecutively. However, this higher coherence score in NVCTM comes with a lower classification accuracy compared to both SCAT and CSCAT, as explained in the previous subsection in addition to lower performance at the document visualization task, as we explain in the following subsection. Overall, CSCAT achieves the second-best coherence score results among the results of the models.

$$npmi(N) = \sum_{j=2}^N \sum_{i=1}^{j-1} \frac{\frac{\log P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)} \quad (3.6)$$

3.4.3 Qualitative Analysis

In this section, we illustrate that our models can learn semantically meaningful representations from textual data. We compare our results to the baseline models, including LDA, k-Sparse, and KATE and ProdLDA using the 20 Newsgroups dataset, with the number of topics set equal to 25. The results are listed in figure 8, 9 and 10 for religion, politics and sports.

We can observe from the figures that our CSCAT model generates the most semantically meaningful topics. Here, we show three topics. The top 10 words learned from the *Religion* category: “resurrection”, “doctrine”, “scripture”, “testament”, “holy”, “jesus”,

“spirit”, “christ” and “faith” are strongly related to Religion. CSCAT also learns meaningful representation for the *Sport* category, including words like “players”, “baseball”, “playoffs”, “leafs”, “scoring”, “league”, and “scored” and under *Politics* topic words like “congress”, “senate”, “clinton”, “president”, “secretary”, “administration” which represent the most meaningful representations among the rest of the words generated by other models.

3.5 Conclusions

CSCAT is a new autoencoder for textual data that we suggested. The models are based on the concept of k -competitive learning, in which only a k number of neurons engage in the learning process while the rest are deactivated. The winning neurons become highly specialized in learning specific properties as a result of the competition. Unlike prior techniques, which introduced competition between the strongest positive and negative neurons, our method removes extremely incoherent neurons first, then adds a competition for the highest and lowest positive and negative neurons in the autoencoder’s bottleneck layer.

Our investigations showed that our method delivers very close or higher performance on a variety of textual data applications, such as classification and topic modeling. Furthermore, compared to the baseline models we examined in this article, our model gives more semantically meaningful topics. Our approach can also be used to reduce the dimensionality of textual data.

Algorithm 5: Approach of Training Phase

```
1 procedure Training Phase:
2   for  $e$  in epochs do
3      $z = \tanh(Wx + b)$ 
4      $H = \text{cscat\_layer}(z)$ 
5      $H = \text{scat\_layer}(k, H)$ 
6      $\hat{z} = \text{power\_aggregation}(z, H)$ 
7      $\hat{x} = \text{sigmoid}(W^T \hat{z} + c)$ 
8      $\text{loss} = \text{criterion}(x, \hat{x})$ 
9      $\text{back\_propagation}(W, W^T, \text{loss})$ 
10  end
11 function  $\text{cscat\_layer}(z)$ :
12   for each neuron in  $z$  do
13      $H \leftarrow \{ \text{neuron} \mid \text{npmi}(\text{neuron}) > \theta \}$ 
14   end
15 S
16 function  $\text{scat\_layer}(k, H)$ :
17    $s_p = \text{get\_strongest\_positive}(k/4, H)$ 
18    $s_n = \text{get\_strongest\_negative}(k/4, H)$ 
19    $w_p = \text{get\_weakest\_positive}(k/4, H)$ 
20    $w_n = \text{get\_weakest\_negative}(k/4, H)$ 
21    $H \leftarrow [s_p, s_n, w_p, w_n]$ 
22 return  $z$ 
23 function  $\text{power\_aggregation}(z, H)$ :
24   for each neuron  $\in z$  and  $\notin H$  do
25      $\text{total\_energy} += E(\text{neuron})$ 
26      $E(\text{neuron}) = 0$ 
27   end
28   for each neuron  $\in H$  do
29      $E(\text{neuron}) += \text{total\_energy}$ 
30   end
31 S
```

Table 9: datasets

dataset	20news	Reuters	Wiki10+
train size	11314	554414	19972
test size	7532	250000	1972
valid size	1000	10000	1000
vocab size	2000	5000	2000

Table 10: Important Notations

Notation	Description
m	Number of Dimension (Topics)
k	Number of winner neurons
z_s	Set of strongest activations
z_w	Set of weakest activations
n	Number of highest activations per topic
E	Energy of the activations

Table 11: Document Classification Results

Model	20 Newsgroups			Wiki10+			Reuters		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
LDA [33]	0.42	0.50	0.46	0.72	0.45	0.56	0.70	0.49	0.44
K-Sparse [27]	0.42	0.42	0.42	0.72	0.45	0.56	0.80	0.50	0.62
NVCTM [35]	0.57	0.56	0.57	-	-	-	-	-	-
KATE [21]	0.70	0.70	0.70	0.73	0.45	0.56	0.70	0.49	0.61
ProdLDA [21]	0.53	0.53	0.53	0.49	0.45	0.47	0.51	0.56	0.47
CSCAT (ours)	0.72	0.72	0.72	0.63	0.58	0.60	0.80	0.52	0.63

Table 12: Coherence Score Evaluation Results

Model	T = 25	T = 50
LDA [33]	0.112	0.140
K-Sparse [27]	0.093	0.090
NVCTM [35]	0.180	0.176
KATE [21]	0.073	0.101
ProdLDA [21]	0.251	0.240
CSCAT (ours)	0.151	0.118

CHAPTER 4

PERSONALITY TRAIT IDENTIFICATION: CHALLENGES AND OBSTACLES

4.1 Introduction

The large amount of text data generated by users on social media has sparked considerable interest in computational personality detection [73], [74]. However, detecting personality using social media writing is a complex issue that requires an understanding of human communication, including social media self-presentation, communication modes, and communication affordance.

Social media allow people to form groups based upon ideology and other human interests [75], [76]. Thus, when one is posting on social media, they target a group of people who share similar interests [75]. There are two defining dimensions of social media: social presence and self-presentation. Social presence reflects technical affordance, for example, multi-media, multi-module, and multi-directional communication [76]. Self-presentation is driven by a human desire to create an image of the "self" for various purposes of social interaction. Studies showed that people might intentionally practice different image management strategies to present others on social media [77] which is different from the self-disclosure in stream-of-consciousness writing (stream-of-consciousness writing used in the Personality Recognizer tool). Individuals with different personalities use different linguistic devices for self-presentation to create a certain kind of image [78]. Thus, personality detection models for social media essentially detect personality types

that an author intends to present to their audience on social media.

This part focuses on human expression and psycho-linguistic patterns and raises three important questions that will be addressed in this study for personality detection. A detailed discussion can be found in Related Research. The main research questions are as follows:

- *RQ1: What embedding methods can produce the most accurate personality detection model for social media texts?*
- *RQ2: Can human psycho-linguistic categorization such as LIWC improve the performance of the model?*
- *RQ3: Are machine learning-based models create more reasonable results for personality detection on social media than Personality Recognizer, which is a tool designed based on stream-of-consciousness writing?*

In this research, we propose a deep learning model for personality detection. Further, we introduce a new approach to analyze the result from the machine learning-based model and PR tool to examine which method would be better for labeling users' personalities in social media. Our experimentation used two datasets; Facebook myPersonality [79], and Twitter data we collected for this study.

4.2 Research Design

4.2.1 Deep Learning Approach and Feature Selection

According to psycho-linguistic research, people of different personality types may tend to use different wording in writing [78], [80]. In other words, specific linguistic markers are indicative of personality types [78], [80].

Our **first aim** is to experiment with various embedding methods of word, sentence, and document and feature selection techniques. These feature extracting and embedding methods have their own strengths and characteristics in discovering the corpus's underlying latent features and patterns. Bidirectional Encoder Representations (BERT) [81] heavily relies on a pre-trained language model, but constrained logic of sequence to sequence pairs [82]. BERT is pre-trained on more extended essays rather than short texts.

Recent efforts on personality detection have been fueled up. [83] proposed a multilayer perceptron (MLP) based on BERT language models combined with psycho-linguistic features. They have used MBTI and essay-based datasets to measure the performance of their models. They achieved better results using Bert model+MLP as opposed to using the psycho-linguistic features (our results confirm the same result). [84] presented transfer learning for personality detection. Their studies explored the efficacy of transfer learning in personality detection. They built a model on top of my personality dataset and used the model to predict the labels of the tweets of 26 users in Twitter.

[85] add an MLP over handcrafted features. [86] proposed a pipeline using deep learning models based on the LSTM and CNN. In this research study, researchers used

different features, including LIWC, SPLICE(Structured Programming for Linguistic Cue Extraction), and SNA (Social Network Analysis).

In another work, a deep learning-based method was proposed [80], [87], where convolutional neural networks are applied to extract n-gram information from stream-of-consciousness essays. Researchers [88] combined author information with features extracted from CNN and fed the combined feature space to the softmax layer. However, these works have not fully realized the potential of the most suitable feature extraction and text embedding techniques that can be used for personality detection. Furthermore, They have not studied the potential difference between personality detection on social media and essay datasets.

Our study explores primary linguistic markers at word and sentence level, using different embedding techniques to detect social media personality. For the measures of the model performance, we reported F1 scores. Also, we explored the efficacy of the machine learning-based model against the traditional psychological tools like PR.

4.2.2 Personality Detection and LIWC

Our **second aim** is to study the efficacy of LIWC [89] on the model building. It theorizes that word-level linguistic markers are indicative of personality types. LIWC developed in the 1990s [90] has inspired efforts of identifying psychometric properties of personal language usage styles. LIWC offers subjective dictionaries that categorize words into standard linguistic dimensions, psychological processes, particular concerns, and spoken categories [90]. These categorizations are not mutually exclusive.

Earlier findings from LIWC showed that standard linguistic aspects and the psychological processes were most helpful in personality detection [90]. In particular, the mental processes offered insights into the choices of words. There was a moderate correlation between the use of social concepts, positive and negative emotion words, and words belonging to the cognitive processes with the five personality categories [90].

Following this research line, this study experimented and compared different feature extraction (TFIDF/Doc2Vec) and feature selection (univariate statistical test; χ^2) methods using machine learning models with LIWC psychological processes. Instead of factoring in individually tailored information, we aimed at extracting maximal information from the text to improve the model performance. We created a vector that is a combination of LIWC and feature/word embeddings. We employed statistic tests to choose the most relevant features with regard to each personality. Our approach is similar to Golbeck's work [91], but they included the highly correlated features in the tweets along with LIWC and MRC Psycholinguistic Database.

4.2.3 Machine Learning vs. Personality Recognizer

The third aim is to design an experiment to check which labeling methods are more suitable for the social media texts; machine learning-based model or Personality Recognizer tool. Thus, we used our proposed classifier in the source domain (Facebook) and demonstrated the transfer learning from the source domain (Facebook) to the target domain (Twitter) so that we can compare the labeling in both Facebook and Twitter.

Transfer learning was introduced to overcome the lack of human-labeled training

data for a classification task in a new domain (called target domain) [92]. The goal of transfer learning is to enhance the performance of learning in a target (unknown) domain by transferring the knowledge learned from the source (known) domain [93]. Thus, transfer learning performance relies heavily on the availability of appropriate auxiliary data of the source domain. The supplementary data can be selected by human intervention or guided by evidence from empirical studies [92]. Recent works [94] demonstrated the effectiveness of knowledge distillation using a model learned from a "teacher" model of the source domain to build a "student" model of the target domain. The student model is learned from the teacher's outputs (called soft labeling) [95] that is required to validate the results with a minimum of human intervention or further refine the model through feedback. Further, researchers [83] studied the use of transfer learning in personality detection. They used the essay and Kaggle dataset to evaluate their proposed model. Their proposed model (Bert+MLP) achieved a better result than the state-of-the-art model on essay and kaggle datasets. Kaggle and Essay datasets have long sentences. One reason that the BERT model does not perform well on the myPersonality and Twitter dataset could be that these datasets have short text; however, Kaggle and Essay dataset contain long texts.

As discussed before, Facebook writing is a many-to-many media mediated communication, while PR [74], built upon stream-of-consciousness writing, which is a monologue. Similar to Facebook writing, Twitter writing is also a many-to-many media-mediated communication and a self-presentation.

According to media theories, a personality type exhibited on social media, like

Facebook and Twitter, is a result of self-presentation (self-image management) [78]. Thus, Facebook myPersonality and PR [74] have fundamental differences in terms of human communication. Facebook myPersonality describes self-presentation of personalities on interactive social media, while PR (the writing data) describes monologue of self-expressive writing. Thus, we hypothesize that transfer learning from Facebook to Twitter would be more effective than the PR method. We even applied PR on the Facebook dataset to check whether labeling with PR tool matches the original labeling. Both Facebook and Twitter results supported the hypothesis.

4.2.4 Evaluating Construct Validity

To answer the third question, "Construct validity" is used. Construct Validity measures whether an operationalization (the method/model) measures what it claims to measure or the approximation of operationalization to the theoretical construct on which the operationalization is based [96]. In our case, the operationalization of personality types is the personality detection outputs from our model. The theoretical construct is social media self-presentation of personality types. Measuring construct validity means we need to evaluate the degree to which our model and prediction outputs are legitimate inferences of the theories about social media self-presentation of personality types. As personality types are latent psychological concepts (in contrast to more observable human traits, such as height and weight), construct validity, meaning whether our model measures what it is supposed to measure, is crucial for social media personality detection. Thus, we employed topic-sentiment cross-categorization and human expert knowledge to

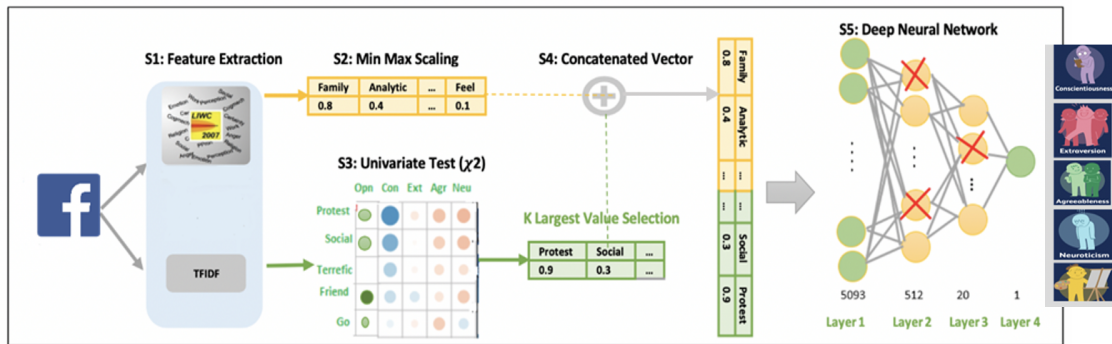


Figure 11: The proposed model (DNN-TFIDF- χ^2)

evaluate the construct validity of the models.

4.3 Research Methodology

4.3.1 Embedding Methods for Personality Detection

In this section, we describe embedding methods and the construction of personality detection models.

We propose a new deep learning (DL) model for personality detection. The process includes five steps; 1) data collection (Twitter data) and pre-processing; 2) feature extraction (TFIDF/Doc2Vec/BERT); 3) feature selection using the univariate statistical test (χ^2); 4) concatenating the vector of the previous step with LIWC dimensions after normalizing the LIWC vector (optional); 5) feeding the vector to the deep learning model. Figure 11 shows the proposed model. The details of the proposed model are explained in Algorithm 1.

Our setting of experimentation is to use TFIDF- χ^2 with and without LIWC. Let us

consider the TFIDF vector result as $T = \{t_i, t_{i+1}, \dots, t_n\}$, where t_i is the value associated with the i feature in the text. We apply χ^2 statistic on each generated feature t_i of TFIDF and target Y_i . Each target, i.e., OPN will be considered an independent vector while computing the univariate feature selection. The univariate feature selection chooses the best features using the univariate statistical test. A small value means that a feature is independent of a label; a large value means the feature is not randomly related to Y_i . The details have been described in Algorithm 1. We choose K of the most correlated features from T . We name this vector T_k . For the predictive models, we have designed a Deep Neural Network (DNN) model that is a four-layer model (as shown in Figure 11).

The proposed model is defined with the DNN-based classifier based on the TFIDF- χ^2 feature vector.

To answer Research Question 2, we added LIWC's predetermined categories on the dataset as part of our experimentation to the TFIDF vector. We represent each sample of data in the dataset as a vector; $L = \{l_i, l_{i+1}, \dots, l_n\}$ in which l_i is a category in the LIWC. We then normalize this vector L_i by min-max scaling, as indicated in line 8 of Algorithm 1.

Finally, The TFIDF- χ^2 features T_k will be concatenated with the LIWC vector L .

In the second setting, we used Doc2Vec. We repeat the same process using Doc2Vec embedding instead of using TFIDF- χ^2 features. In the Doc2Vec embedding the dimensions have already been reduced. Thus the univariate statistic test was not applied to select among the dimensions.

Algorithm 6: Embedding features

```
1 function select best features( $L, T, Y$ ):  
    // compute the univariate test w/t TFIDF values &  
    target  
2  $observed = Y \cdot T$   
3  $feature = \sum_{i \in T} T_i$   
    // c is the number of distinct classes; 2  
4  $prob = \sum_{i \in c} Y_i / |c|$   
5  $expected = prob \cdot feature$   
6  $chi = \sum_{i \in T} \frac{(observed - expected)^2}{expected}$   
7  $T_k = [\text{largest } k \text{ values in } chi]$   
8  $L_s = \frac{c_i - \min_{v \in L} c_i}{\max_{v \in L} c_i - \min_{v \in L} c_i}$  // concatenate two vectors  
9  $vec = [T_k \cdot L_s]$   
10 return  $vec$ 
```

4.3.2 Validation Method: Topic-Sentiment Cross-Categorization

We designed a cross-categorization topic-sentiment to check which labeling methods are more meaningful and closest to the theoretical principles. Unlike existing evaluation methods that aim at accuracy, our validation check's goal is to examine construct validity: which labeling method (machine learning-based model or PR) can detect personality types in line with psychological and media research theories. As discussed earlier, we hypothesize that our proposed model is a more valid/appropriate labeling method than PR.

PR takes a user's corpus and computes estimates of personality scores along with the Big Five model. PR is built based on the LIWC categories and additional dictionaries, which categorize words into standard linguistic dimensions, psychological processes, personal concerns, and spoken classes [90]. We compared the ground-based data labeled by

PR and Facebook myPersonality, to evaluate which labeling method, using our proposed model, is more suitable for personality detection on Twitter. First, KATE topic modeling [21] is used to extract the most characteristic features of the textual data. KATE is a k-competitive autoencoder model that aims to extract topics from the documents. We attempt to understand what percentage of the topics retrieved from the text data are covered and related to the LIWC categories.

Second, we have extended the well-established sentiment analysis tool VADER [97] into topic-wise sentiment analysis to generate evidence for personality detection’s construct validity. Initially, we classified the Twitter and Facebook datasets in terms of the LIWC categories. Then, we have applied topic modeling to each personality trait. Each personality topic, generated from the existing Facebook myPersonality/Twitter dataset, was mapped to the LIWC category annotation. We calculated the percentage of the topics covered in the LIWC categories in each sentiment group (positive, negative, neutral). In other words, we computed the topic of specific sentiment weight that was covered in the LIWC categories. Equation 4.1 describes the topic-sentiment categorization (sc). Later, we review how this topic-sentiment cross categorization can explain the theoretical underpinning of personality and self-presentation.

$$sc(w) = \sum_{w \in c} \frac{1}{count(w)} * 100, \text{ for } i \in (\alpha, \delta, \gamma), \text{ if } count(w) > 0 \quad (4.1)$$

where c is LIWC categories, w a word in topics, α, δ, γ are sentiment categories.

4.4 Experimental Results

Table 13: Dataset for Experiments

Facebook Dataset		Twitter Dataset	
#Users	250	#Users	29
#Reports	9,917	#Reports	258,301
#Reports (Filtering)	7,687	#Reports (Filtering)	218,629
#Words	146,159	#Words	5,302,073
#Words (NLP)	64,187	#Words (NLP)	1,968,660
#Total Reports	38,435	#Under-Sampling Reports	326,428
#Reports (O)	Y:5700 N:1987	#Reports (O)	Y: 5208 N: 5208
#Reports (C)	Y:3493 N:4194	#Reports (C)	Y: 54307 N: 54307
#Reports (E)	Y:3289 N:4398	#Reports (E)	Y: 21753 N: 21753
#Reports (A)	Y:4119 N:3568	#Reports (A)	Y: 67325 N: 67325
#Reports (N)	Y:2870 N:4817	#Reports (N)	Y: 14621 N: 14621

4.4.1 Experimental Setting

The experimental setting uses a single NVIDIA GeForce RTX 2080 Ti GPU board (11GB), i9 processor (8 Cores) and RAM of 32 GB. Facebook myPersonality [79] relies on 180,000 self-reported personality surveys [98] to validate personality types for Facebook writing. PR [74] uses two datasets and two validation methods. The first set of data is 2,479 essays from psychology students' stream of consciousness writing [90]. The Big Five Inventory questionnaire assesses the participant's personality. The second set of data is the Electronically Activated Recorder (EAR) corpus [99] consisting of transcripts of conversational partners, whose personality is evaluated by both self-reported surveys and

observers' evaluation [74]. We used Natural Language Processing (NLP) techniques, including removing stopwords, performing text normalization, fixing punctual errors. The original data of the Facebook Big5 dataset is multi-labeled. We have converted the Facebook Big5 to the binary labeled data for the five personality classes. We have two classes (negative 0 and positive 1) for each category (O, C, E, A, N) in the personality dataset. The imbalance problem exists with the openness class of the Twitter dataset. The initial dataset had an imbalance ratio of 3% vs. 97% (positive vs. negative), or <3% positive samples were contained in this dataset. We used the random undersampling approach that randomly deleted samples in the majority class to overcome this problem. To build the twitter models, we have used Facebook's myPersonality dataset [100]. Table 13 shows the details of the datasets we used in our experiments.

4.4.2 Baseline Models

Table 14 shows the details of the feature vectors used by configurations of two deep learning models, including deep neural network and BERT. The results of our model are compared to the following baseline models:

- Bidirectional Encoder Representations from Transformers (BERT) based binary classifiers: We have built three personality detection classifiers using the Bidirectional Encoder Representations from Transformers (BERT) [81]. The hyperparameters for the BERT fine-tuning models were defined as follows: learning rate of 1e-5, epoch of 20, optimizer of ADAM 1e-8, max sequence length of 100, the batch sizes of 8. The binary classifiers are designed with the sigmoid function for personality

Table 14: Overall Experiment Settings

Feature & Embedding Vectors			
Min Frequency of Word	5	Document Dimension	50
#TFIDF Features after Chi-2	2,000	LIWC Dimension	93
TFIDF Embedding Dimension	2,093	Doc2Vec Embedding Dimension	50
BERT Embedding Dimension	512	BERT* Embedding Dimension	512
Machine Learning			
Data Split:	Training 70%	Testing 20%	Validation 10%
<i>BERT Classifier</i>		<i>Deep Learning Classifier</i>	
BERT #Epoch	20	DNN #Epoch	10
#Transformers	12	#Epoch Doc2vec	50
#Layers per Transformers	11	#Node First Layer	100
#Layers for Custom Classification	6	#Node Second Layer	20
BERT Activation Function	Tanh	Activation Function of DNN	Tanh
Last Layer Activation Function	Sigmoid	Last Layer Activation Function	Sigmoid
BERT Dropout	0.3	DNN Dropout	0.5

detection using the following three fine-tuned BERT models:

1. Pretrained BERT + binary classifier
2. Retraining the Bert on Facebook dataset + binary classifier
3. Retraining BERT + LIWC + binary classifier

For the retraining with Facebook and Twitter datasets, the BERT* architecture consists of (i) 12 x transformers, based on the Base Uncased version. (ii) Custom Classification layers including Dropout (with a value of 0.3), three hidden layers of rectified linear units (ReLU) (768x512, 512x256, 256x2). For the optimization, we used Adam optimizer and BCEWithLogits for Loss Function. For the BERT*, the architecture was revised for taking the input of Facebook and Twitter datasets

together with LIWC features.

- SVM + LIWC + Social Network Analysis + Time related features [101]: In this work, first, multiple features were extracted from the text data, including 81 features of the LIWC categories, seven features of social network (network size, density, and transitivity), time-related features (frequency of status updates per day, number of statuses posted between 6-11 am, number of statuses posted between 11-16), and other features (total number of statuses per user, number of capitalized words). Second, they combined all these features and passed them to different classifiers. The SVM classifier with LIWC+Social+Time showed the best performance among multiple classifiers.
- CNN [88]: They designed two models; one works based on Recurrent Neural Network and another one based on CNN. The CNN model consists of three layers in which the first layer is an embedding layer; the second layer is a convolution layer on top of the embedding layer to extract unigram/bigram/trigram features, the third layer is to concatenate these features and pass them to the fully connected layer for the classification.
- Semi-supervised approach [102]: In their research, to take advantage of huge unlabeled data, they adopted Pseudo Multi-view Co-training (PMC) [103], an effective Semi-supervised learning algorithm, to build a personality prediction model. To extract adequate linguistic features of LIWC and n-gram, they used a method based on word embedding [104] to avoid the sparsity of word-based features.
- [105] presented three classical models for personality detection (BERT Single,

CNN Single, LSTM Single). From their work, the CNN Single achieved the best performance compared to the other two personality trait detection models due to CNN’s feature extraction ability.

4.4.3 Personality Detection Classification Results

4.4.3.1 Comparative Evaluation

We have conducted a comparative evaluation of personality detection with our proposed model and state-of-the-art approaches in several different settings.

The F1 score comparison of the baseline models was based on their published results. First, the proposed model is superior to these approaches for personality detection on the Facebook dataset with/without LIWC (Table 15). Specifically, the Deep Neural Network and TFIDF- χ^2 (DNN-TFIDF- χ^2) is the best in Facebook with LIWC embedding, and the DNN and Doc2Vec are the best in Facebook without LIWC embedding. We can see that our proposed model outperformed the baseline for all the available models.

Interestingly, the result of our study indicates that in the BERT model on the Facebook datasets with/without LIWC, ”openness” achieved higher accuracy compared with other personality traits.

As shown in Table 15, it is clear that LIWC, in most cases, did not introduce a significant increase in the model’s performance. Notably, when LIWC is added to Doc2Vec, the performance drops significantly. It seems that machine learning-based feature extraction works better than the more controlled embedding method (deductive methods) to predict personality based upon social media text. As noted earlier, LIWC categorization

Table 15: Facebook Binary Classification for Personality Detection (F1 Score%)

Approach	Facebook with LIWC						Facebook without LIWC					
	O	C	E	A	N	AVG	O	C	E	A	N	AVG
DNN-TFIDF- χ^2 (ours)	80.4	76.0	74.6	76.1	74.4	76.3	81.2	77.5	75.9	78.5	76	77.8
DNN-Doc2Vec (ours)	70.0	52.0	52.6	47.5	50.1	54.4	81.5	86.0	89.3	84.3	77.3	83.7
BERT*-Sigmoid (ours)	87.0	45.0	52.0	71.0	26.0	56.2	87.0	59.0	43.0	62.0	44.0	59.0
BERT-Sigmoid	73.4	56.3	58.1	53.4	63.3	60.9	73.4	56.3	58.1	53.4	63.3	60.9
SVM [86]	63.0	54.0	60.0	58.0	57.0	58.4	–	–	–	–	–	–
CNN [88]	–	–	–	–	–	–	61.3	53.3	65.2	57.0	62.7	59.9
PMC [80]	65.0	62.0	71.0	68.0	70.0	67.2	–	–	–	–	–	–
CNN Single [105]	–	–	–	–	–	71.5	–	–	–	–	–	–
LSTM Single [105]	–	–	–	–	–	69.1	–	–	–	–	–	–
BERT Single [105]	–	–	–	–	–	68.4	–	–	–	–	–	–

BERT*-Sigmoid: Our binary classification with BERT retraining with Facebook and/or LIWC
 The results by others were based on the published ones. Some of them specified by the symbol
 "–" were not available.

is predetermined and not mutually exclusive. Thus, LIWC contribution to classification will require more excellent coverage across all personality categories to produce sufficient discriminability. However, such ideal LIWC distribution does not often exist in text data since people tend to intentionally use certain wording to achieve interpersonal and media-mediated communication goals.

4.4.3.2 Twitter Model Results

The Twitter model was built through transfer learning using the proposed model on the Facebook model. Besides the validation of the Twitter model with Personality Recognizer to be discussed later, we have conducted a similar comparative evaluation for the Twitter dataset with the the DNN-TFIDF- χ^2 individual models.

Table 16 shows the binary classification performance of the four different approaches on the Twitter dataset. Similar to Facebook’s personality detection results,

the DNN-TFIDF- χ^2 classifier showed the best performance for Twitter/LIWC with an average accuracy of 85.8%. The DNN-TFIDF- χ^2 classifier shows the best overall performance (the overall average accuracy of 83.85% and 82.7%) for the Twitter data both with/without LIWC.

Table 16: Our proposed model’s Accuracy for Twitter Personality Detection (F1 Score%)

Approach	Twitter with LIWC						Twitter without LIWC					
	O	C	E	A	N	AVG	O	C	E	A	N	AVG
DNN-TFIDF- χ^2	92.7	82.5	85.0	81.4	87.4	85.8	62.1	82.0	84.7	81.2	87.8	79.6
DNN-Doc2Vec	79.9	88.9	89.6	88.4	76.2	84.6	64.5	85.3	86.0	86.3	67.7	78.0
BERT*-Sigmoid	86.0	77.0	83.0	69.0	84.0	79.8	78.0	77.0	84.0	73.0	79.0	78.2
BERT-Sigmoid	86.0	72.0	83.0	69.0	84.0	78.8	78.0	77.0	84.0	73.0	79.0	78.2

The DNN-TFIDF- χ^2 model achieves a higher F1 accuracy on both Twitter and Facebook data. Under two different settings, TFIDF- χ^2 outperforms Doc2Vec, as is shown in Table 16. Achieving better results of available models suggest that a combination of feature extraction, χ^2 statistic tests to find the most relevant features for predicting each personality trait can improve results significantly. On the other hand, Doc2Vec/LIWC did not obtain a comparable result. It may suggest that frequency-based feature extraction would better indicate personality than semantic-based feature embedding for personality detection. Another finding is that the accuracy of the model to predict Neuroticism in Twitter is higher than Facebook. This finding suggests that language use on Twitter may exhibit a more towering personality of Neuroticism than that on Facebook, which could be attributed to polarization in the Twitter discussion.

Table 17: Facebook Topic-Sentiment Cross-Categorization: DNN-TFIDF- χ^2 Classifier vs. Personality Recognizer

Cat	Sen	Facebook Dataset											
		DNN-TFIDF- χ^2 Facebook Classifier						Personality Recognizer					
		Soc	Bio	Aff	Cog	Rel	Per	Soc	Bio	Aff	Cog	Rel	Per
O	POS	83.9	84.9	81.6	53.9	33.7	63.3	88.2	78.9	74.7	47.0	32.6	53.8
	NEG	5.3	9.4	16.4	16.9	18.0	3.0	11.7	21.0	23.4	29.4	34.7	46.1
	NEU	10.7	5.6	1.8	29.1	48.1	6.6	0.0	0.0	1.8	23.5	32.6	0.0
C	POS	71.4	59.6	72.3	49.4	36.1	48.3	75.0	26.6	79.7	51.5	37.0	62.9
	NEG	12.2	26.9	25.7	21.9	11.1	12.9	4.16	53.3	18.9	12.1	11.1	3.7
	NEU	16.3	13.4	1.9	28.5	52.7	38.7	20.8	20.0	1.26	36.3	51.8	33.3
E	POS	86.8	80.0	85.6	41.1	37.1	50.0	94.5	88.2	86.4	41.8	33.8	42.8
	NEG	13.1	10.0	13.7	24.7	25.6	8.3	0.0	2.94	13.5	9.3	13.2	14.2
	NEU	0.00	10.0	0.6	34.1	37.1	41.6	5.40	8.82	0.0	48.8	52.9	42.8
A	POS	71.6	59.4	82.4	53.3	34.4	69.5	33.3	22.2	85.7	39.2	22.7	100
	NEG	10.0	33.3	16.5	16.9	21.6	17.3	33.3	66.6	7.14	21.5	22.7	0.0
	NEU	18.3	7.2	1.0	29.6	44.0	13.0	33.3	11.1	7.14	39.2	54.5	0.0
N	POS	74.1	62.5	68.2	56.0	28.1	69.2	0.0	0.0	0.0	0.0	0.0	9.09
	NEG	9.60	35.0	30.2	12.1	22.9	15.3	0.0	0.0	100	50.0	50.0	0.0
	NEU	16.1	2.50	1.55	31.7	48.9	15.3	0.0	0.0	0.0	50.0	50.0	0.0

Soc: Social, Bio: Biological, Aff: Affective, Cog: Cognitive, Rel: Relative, Per: Perception
O: Openness, C: Conscientiousness, E: Extraversion, A: Agreeableness, N: Neuroticism

4.4.4 Construct Validity

The post-processing analysis serves two purposes: 1) comparing two labeling methods, DNN-TFIDF- χ^2 vs. PR; 2) assessing construct validity of the models (DNN-TFIDF- χ^2 vs. PR). To achieve these two goals, this study designs topic-sentiment cross-categorization as explained in Equation (1). Table 17 summarizes Facebook and Twitter topic-sentiment percentage from LIWC’s categories. A human expert conducted an interpretation and validity check.

Table 18: Twitter Topic-Sentiment Cross-Categorization: DNN-TFIDF- χ^2 Classifier vs. Personality Recognizer

Cat	Sen	Twitter Dataset											
		DNN-TFIDF- χ^2 Twitter Classifier						Personality Recognizer					
		Soc	Bio	Aff	Cog	Rel	Per	Soc	Bio	Aff	Cog	Rel	Per
O	POS	43.2	72.9	86.6	48.0	32.4	27.0	56.1	69.5	79.8	43.9	26.3	35.1
	NEG	27.6	9.3	12.1	22.1	31.7	24.5	35.6	13.5	18.7	29.0	47.8	36.1
	NEU	29.0	17.6	1.1	29.7	35.8	48.3	8.1	16.8	1.37	27.0	25.8	27.8
C	POS	42.4	72.6	83.2	45.8	27.6	34.5	37.2	55.2	90.9	48.4	38.0	28.3
	NEG	28.1	6.80	15.2	25.2	36.4	16.7	28.2	9.93	8.04	19.7	23.6	13.4
	NEU	29.4	20.4	1.5	28.9	35.8	48.6	34.5	34.7	1.04	31.8	38.2	57.8
E	POS	40.7	62.9	84.5	49.1	33.3	28.0	50.7	85.9	91.4	51.9	36.0	25.8
	NEG	33.5	14.9	13.6	26.6	32.5	22.5	24.0	3.92	7.74	15.6	24.9	22.3
	NEU	25.6	22.1	10.8	24.1	34.1	49.3	25.1	1.0	0.8	48.8	39.0	51.7
A	POS	45.7	82.1	86.9	53.1	35.6	36.2	39.9	60.4	87.3	50.3	30.7	22.9
	NEG	27.2	5.8	11.7	20.9	22.9	23.1	28.8	11.6	11.8	16.5	18.7	28.4
	NEU	27.0	11.9	1.2	25.8	41.3	40.6	31.2	27.0	0.7	33.0	50.5	48.6
N	POS	24.8	27.2	70.9	48.0	47.0	41.6	11.9	0.0	12.9	5.07	0.0	9.09
	NEG	33.1	45.4	25.1	25.2	25.0	28.4	56.7	100	85.4	63.7	87.1	66.6
	NEU	41.9	27.2	3.90	26.6	27.9	29.9	31.3	0.0	1.61	31.1	12.8	24.2

Soc: Social, Bio: Biological, Aff: Affective, Cog: Cognitive, Rel: Relative, Per: Perception
O: Openness, C: Conscientiousness, E: Extraversion, A: Agreeableness, N: Neuroticism

4.4.4.1 Assessing Construct Validity

We conclude that DNN-TFIDF- χ^2 produces more valid and interpretable results for social media text than PR. When Twitter datasets are labeled by our model trained on myPersonality Facebook dataset (see Table 17, five personality traits consistently show more positive sentiment about the LIWC social (friends/ family/ humans) process on Facebook than Twitter. This finding aligns with previous studies: Facebook expresses more positive emotion about family and friends [106]. Besides, using this labeling

method, the results also show more positive sentiment about the perceptual (seeing/ hearing/ feeling) process on Facebook than Twitter across all personality types. This finding again supports the social and life-oriented self-presentation of Facebook [106]. However, when labeled by Personality Recognizer (both Facebook and Twitter) [74] in Table 18, the sentiment about the social and the perceptual processes are less consistent primarily due to discrepancies in Neuroticism. Additionally, both labeling methods produce more positive than negative emotional sentiment on Facebook and Twitter, except Neuroticism labeled by PR.

The comparison of machine learning-based model and PR suggests the detection of openness performs consistently when different labeling methods are used. Openness and conscientiousness have slightly more positive cognition on Facebook than Twitter, while extraversion has a more positive perception on Twitter than Facebook (consistent between the two models). These findings suggest that extrovert users like to express positive thinking in conversation (Twitter), and open and conscientious users want to share positive thinking about their lives (Facebook). It seems that openness, as a socially desirable personality trait, is present similarly across various media platforms and communication modes (monologue or conversations). These findings suggest that Twitter and Facebook are used differently for the cognitive or "thinking aloud" process by different personalities. Previous studies show that Twitter focuses more on conversation and Facebook on personal life [106].

The affective process indicates general trends of being emotionally positive in a self-presentation on social media and thus is less sensitive to differences in media

platforms. In contrast, the perceptual and social processes reveal consistent and significant media differences in personality detection. The perceptual and social processes are more positive on Facebook than Twitter, suggesting that users are more likely to use Facebook than Twitter to share positive emotions about family, friends, and life experiences. The biological process demonstrates significant variation with myPersonality and PR. PR shows more negative than positive sentiment about the biological process (body/health/sexuality/ingestion) on Facebook regarding agreeableness and conscientiousness.

A possible explanation is that PR may categorize negative sentiment expressed about others' health and illness into the authors' personality markers. Relativity also varies with the two different models, without any consistent patterns. These findings suggest that relativity may not be a reliable predictive aspect of personality detection.

considering these factors, this study summarizes the self-presentation of personality on Facebook and Twitter as follows: Facebook is geared toward an image of a positive personality who is open, conscientious, extravert, and agreeable. Twitter, assuming that it is more opinionated, is reflecting more negativity due to possible polarization in the discussion. Here discussion should focus more on the media difference in twitter's debate.

4.5 Discussion

One of the exciting findings from this study is that the frequency-based feature-based approach is better than semantic-based feature embedding for personality detection.

Further, the BERT-based model did not show a consistent performance in terms of accuracy for personality detection. Thus, the DNN-TFIDF- χ^2 model is better than the BERT model. We argue personality detection using social media writing should recognize the influence of media affordance and conceptualize personality as the antecedent of self-presentation.

Second, the machine learning-based model performs better at predicting and labeling social media texts than Personality Recognizer. Further, Facebook shows more positive thinking about self from open and conscientious users, while Twitter shows more positive thinking of extravert users in conversation. As a media platform, Facebook facilitates the presentation of more positive social and perceptual aspects of users' life more than Twitter. The results confirm that Twitter has a different self-presentation pattern of personality from Facebook, and the differences could be attributed to the character and technical affordance.

4.6 Conclusion

This research proposes a deep learning model to predict personality on social media data such as Facebook and Tweets. Our results showed that due to the self-presentation nature of social media, machine-learned features and models tend to perform better than predetermined embedding techniques like LIWC. Further, the proposed model can deliver good performance compared to pre-trained transformer encoders, like BERT. It seems that the style of language used on social media is more pronounced than the content of the text to predict personality. Also, from the perspective of construct validity, we can see that

conceptually, a machine-learned model is more appropriate for labeling social media text than Personality Recognizer, which was developed using monologues.

CHAPTER 5

PUBLIC DISCOURSE ABOUT THE OPIOID CRISIS ON TWITTER, 2010-2019

5.1 Introduction

The mass media is an important public sphere for public discourse on health issues and providing perspectives to shape public health policies [107]. "Public discourses" are different from "private discourses;" they are motivated by the pursuit of the public good rather than seeking private interests or relationships [108]. "Public discourses" about health aim at serving the common good of people's health.

One major public health concern that has entered the public discourse is the opioid crisis. In the past 10 years, the US is experiencing the worst opioid epidemic in history [109]. The opioid overdose deaths were 16,849 in 1999, and the number was tripled in 2015 to 52,404 [109]. To raise public's awareness and gather support to solutions, studies turned to the mass media and examined the coverage of health-oriented solutions [110]. With the increasing popularity of social media, more and more studies recognize the importance of adding social media to the examination of public discourses about health [111].

Social media data is complex since it serves multiple functions of health communication, which can be broadly categorized into private utterances and public discourses [112]. The richness of social media data allows us to use private utterances (content about patients' private states) to track incidents of diseases and obtain insight of patients.

Meanwhile, Kotlier [112] describes online health forums as “sandboxes” for people to form narratives about health. In other words, social media data documents the formation of public discourse on health.

Recent studies indicate that the public discourse on social media, featuring its own “agenda,” operates independently of the news media discourse [113]. It is an equally important data source, contributing to a comprehensive understanding of the public discourse on the opioid crisis. However, very few studies have shed light on the social media public discourse about opioids. To bridge this gap, this study aims to analyze its major themes and sentiment, to deepen the understanding of its formation, evolution, and current state. Our analysis provides public health professionals and policy-makers with a venue of listening to the public for informed decision-making. Our methods illustrate the complexity of social media data and demonstrate ways of fully leveraging it for multiple analytical purposes.

5.2 Literature Review

To research public discourse, often news stories are among the top choices. News media is not only a platform for disseminating health information [114]; it also influences Americans’ support for various solutions to the opioid crisis in the public health policy domain [110], [115]. In recent years, studies in health communication noted that more and more people turned to social media for various purposes relating to health. Studies have noted two roles that social media play in public health [116]: 1) Exchanging informational and psychological support among patients and

their friends and families [117]; 2) constructing a collective understanding of public health issues [118], [119]. The second usage of social media makes it a “new public sphere” for public discourses on health. For example, people use social media platforms to form narratives around illnesses and health [112] [116], and express their concerns and opinions on the opioid crisis [110].

5.2.1 Private Utterances on Social Media

Existing studies analyzing social media often focused on the first role and used this part of the data to understand patients’ experiences and track and monitor cases of infection. Relating to opioid abuse patients, a recent text data-mining analysis of 100 Reddit posts by 73 unique usernames and all the comments to these posts showed that 66 percent of these authors reported severe to mild opioid use disorder (OUD), 43% have sought information, and 24% support. 95% of the comments provided content that contained therapeutic factors, and 67% shared personal stories to encourage the authors [109].

The individual level data from social media has inspired exploration of using supervised machine learning to conduct geolocation-centric monitoring of the US opioid epidemic in near real-time [120]. Due to rich data on individual opioid experiences, Fan et al. [121] proposed AutoDOA, a transductive classification model to automatically detect opioid addicts based upon Tweets. Another example is a transfer-learning model was applied to identify the effective alternative treatment of OUD from a large amount of social media data [122]. Analysis of Twitter data also shows a strong correlation between misuse of prescription opioids discussion and state-by-state estimates of the misuse by

2013-2015 National Surveys on Drug Usage and Health [123]. These types of existing studies demonstrate **social media data can be used to track and monitor infection rates and disease incidences.**

While the tracking studies reported a high correlation with CDC's data on disease incidences [124], a gap in the analysis is to discover patients' social and psychological processes of experiencing the symptoms and the treatments by examining the highly personal writing [90] [125]. A 2017 analysis of 51,537 posts by 16,162 unique users on "subreddit r/opiate" from 2014 to 2017 applied LIWC-based psychological categorization, and LDA topic modeling [126]. The results suggested that people used social media to constantly discuss their usage of opioids, and the majority of the topics were related to individual experiences of opioids, like opioids affecting social life [126]. Thus, **social media data can also be used to obtain an in-depth understanding of the patients**, so effective health education intervention can be designed [116].

5.2.2 Public Discourses on Social Media

The significant gap that this study tries to bridge is examining the public discourse of the opioid crisis on social media. As aforementioned, the second role that social media plays is a public forum for collective discussion on health issues. A few recent studies compared public discourse on social media with news reporting from agenda-setting's perspective [113]. The empirical evidence shows that public discourses on social media do not have a mechanical connection with the news agenda, suggesting two parallel processes that coincide with each other on specific issues [113].

The agenda-setting theory is concerned with the mechanism, or specifically, news media's role, of setting issue priorities for public discourse [127]. The news media, especially legacy media such as the New York Times and the Washington Post, has played an essential role in setting the agenda of the public discourse by influencing the public's prioritizing of "what to think about" [113].

Previous studies on agenda-setting of the opioid crisis showed that news reporting on criminal justice solutions was trending down from 1998 to 2012, and prevention-oriented solutions trended up [128]. By analyzing a random sample of 600 opioid news stories on solutions between 2013 to 2017, McGinty et al. [110] revealed the major themes were treatment, harm reduction, and medication treatment. The significant increase of opioid-related treatments in 2015 to 2016 was also noted by a recent study on national print news between 2007 and 2016 [129].

A standard question that the agenda-setting research attempts to answer is "who sets the agenda?" [113]. The news media can set the agenda since journalists act collectively based upon shared values and practices of the journalism professionalism [113]. For example, studies showed that the journalists' communities might react in similar ways to handle reports from right-wing alternative media outlets to protect the journalism professionalism and institution [130].

Different from the news media, social media is a platform for patients, their friends, and families, and the general public to have public discussions and collectively construct understandings of public health issues. Everyone and anyone can participate in public discourse on social media, including many people who are not indoctrinated in

journalism professionalism. Thus, the mission of upholding journalism professionalism is not part of the social norms or expectations of public discourse on social media. Although news media actively engage in social media discussion as opinion leaders, the mechanism of the traditional sense of “agenda-setting” process does not hold the ground to prevail on social media.

Instead, studies found that social media discussion is a process of collective meaning construction. [118] A stream of studies examined the meaning-transfer process of obesity and overweight in all-peer online health communities [118]. Collective understanding formed on social media reflects local culture regarding weight management [119]. The findings suggest that changing collective understanding of public health issues may help to mitigate the prevalence of obesity and overweight [119].

Similarly, the framing process on social media differs from that of the news media. In public health, framing is as crucial as the agenda-setting process to influence public discourse and policymaking. Frames are mental structures or schemata of organizing and simplifying the world. Using the public health model of reporting, the journalistic community may choose to collectively frame opioids as a health issue or a criminal justice issue. The Cincinnati Enquirer’s Heroin Beat exemplifies this framing process [131], [132]. Their success shows that news media play an important role in promoting medical and health policy solutions and highlighting social determinants of health [131], [132].

On social media, journalists are not the major driving force of framing; ordinary users are. They are both creators and consumers of media frames. Ordinary users collectively choose specific frames to present their reviews on the opioid crisis. Especially

regarding the sentiment of various aspects of the issue, frames of social media can provide additional insight to policymakers. Therefore, **social media is an important source to research public discourses about health issues.** In this study, our analysis focused on the public discourse of the opioid crisis on social media over time. Meanwhile, we demonstrated that different parts of social media data could be analyzed to provide multiple types of intelligence.

5.3 Data And Methods

This study conducted a natural language processing (NLP) analysis of opioid-related tweets from 2010 to 2019 (see Table 20). We also compared themes of tweets with those of the news reports from NYTimes.com (see Table 19).

5.3.1 Study Sample

A total of 162,760 tweets of “opioid,” “opioids,” and “opioidcrisis” and 2,998 news stories from January 2010 to December 2019 were scrapped from Twitter and NYTimes.com. New York Times is among the highest-circulation national newspapers in the U.S. [110].

5.3.2 Data Analysis

First, text data preprocessing was conducted on both tweets and news stories. Stopwords, punctuation, UTF characters, emoticons, URLs, and hashtags were removed from tweets. All text was converted into lower cases.

Table 19: Opioid News Articles from the New York Times 2010-2019

year	Number of news
2010	6
2011	14
2012	16
2013	21
2014	40
2015	46
2016	251
2017	680
2018	942
2019	892

Second, two subsets of tweets were created. The first subset included “public discussion tweets.” These tweets were part of the public discourse about the opioid crisis, no matter whether the author was directly impacted or not by opioid abuse. For example, “Students across the country want to share stories and want to see coverage about the crisis;” and “Americans who lost live to overdose and drug overdose rate have grown recently.” The second subset included “personal experience tweets.” These tweets were primarily about the person who wrote the tweet. They were often directly impacted by drug abuse, and thus, their writing tended to be highly emotional and/or opinionated. For example, “I hope I can get more help. I hope no pain today.” “Four years ago today, my brother died from drug abuse. I still think of him.”

This division is necessary since the research aims to measure social media public discourse of opioid. Public discourse is different from “private discourse,” which is essentially about writers’ “private states.” It also expects “civility” and excludes “incivility” such as rude comments, hostility, and self-righteousness [108]. In other words, a valid

Table 20: Opioid Tweets from 2010-2019

Year	Total	Personal Experiences	public discourse
2010	159	0	159
2011	708	4	704
2012	2238	4	2234
2013	3676	14	3662
2014	5610	35	5575
2015	11189	14	11175
2016	27298	56	27242
2017	45841	441	45400
2018	33094	2105	30989
2019	32947	1955	30992

input to a public discourse should not be highly personal or highly opinionated. It should include a certain amount of fact-driven discussion and reasonable opinions to serve the “public good.”

We employed a clue-based approach using the Subjectivity Lexicon of MPQA (Multiple Perspective Question Answering) opinion corpus to create such a subtle differentiation. MPQA was created to identify expressions of private states in a sentence [133]. Private states cover opinions, evaluations, emotions, and speculations [133]. MPQA is coded at the sentence level, taking into consideration the context [134]. Its Subjectivity Lexicon differentiates strongly subjective clues and weakly subjective clues. Strongly subjective clues are the ones that are almost exclusively used with subjective meaning, while weakly subjective clues are the ones that can be used to express both subjective and objective meaning [133].

The threshold of determining subjectivity varies from study to study [133]. Wiebe and Riloff (2005) [134] operationalized subjective expressions as ones with one or more

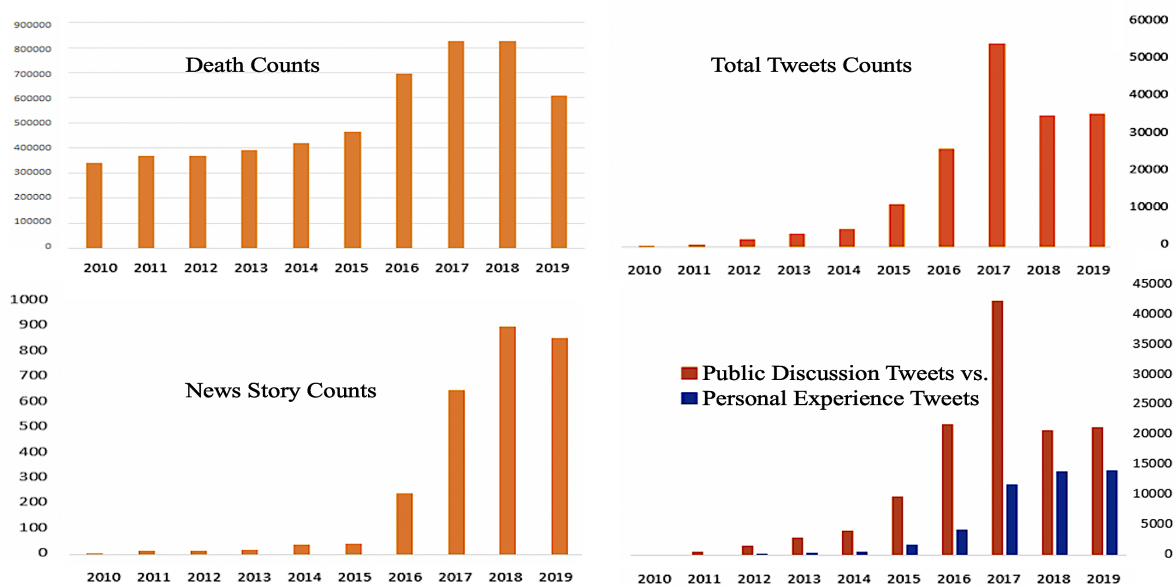


Figure 12: Year-by-year Comparison of Death Counts, News Stories, and Tweets

medium or higher subjectivity clue. Ji et al. (2015) [111] used the threshold of two strong subjective clues and one weak clue, with the assistance of profanity check, for “personal tweets” when analyzing public health concerns expressed on social media. Wiebe and Riloff’s (2005) threshold [134] for subjectivity is lower than that of Ji et al. (2015) [111]. Considering that social media writing is more casual than news writing, we adopted one strongly subjective clue and one weakly subjective clue as the threshold of “personal experience tweets,” with the assistance of a profanity check. We used the same profanity checklist used in Ji et al. (2015) [111]. Figure 12 shows proportions of public discussion tweets and personal experience tweets in comparison to the death counts and news reports, year by year.

The third step was to perform KATE topic modeling on all tweets and news stories to discover main topics year by year. Then, the outputs were separated by types (public

discussion tweets, personal experience tweets, and news stories) and by years. The sentiment analysis with VADER was conducted to analyze the general sentiment about opioid issues in public discussion tweets and news reporting.

The fourth step employed the LIWC (Linguistic Inquiry and Word Count), a lexicon of the psychometrics of word usage (see Figure 21), and VADER sentiment analysis to closely examine the emotions and feelings expressed in personal experience tweets. VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool. The LIWC2007 Dictionary is composed of 2,290 words and word stems. Each word or word-stem defines one or more word categories or sub-dictionaries. For example, the term “crying” belongs to three categories: sadness, negative emotion, and overall affect. Hence, if it is found in the target text, each of these three sub-dictionary scale scores will be incremented.

The complete data collection and analysis process for this study is illustrated in Figure 13.

5.3.2.1 KATE Topic Modeling

KATE (K-competitive Autoencoder for Text) is a deep learning topic model that performs unsupervised text classification. A basic autoencoder is a shallow neural network that consists of two parts: encoder and decoder. The encoder maps the input layer x to the bottleneck space $z = g(Wx + b)$ while the decoder reconstructs the input at $\hat{x} = g(W^Tz + c)$. Here, W^T is the weight matrix obtained by weight tying, also known as weight sharing, i.e., setting $W\hat{a}^2 = W$, which is often used as a regularization method

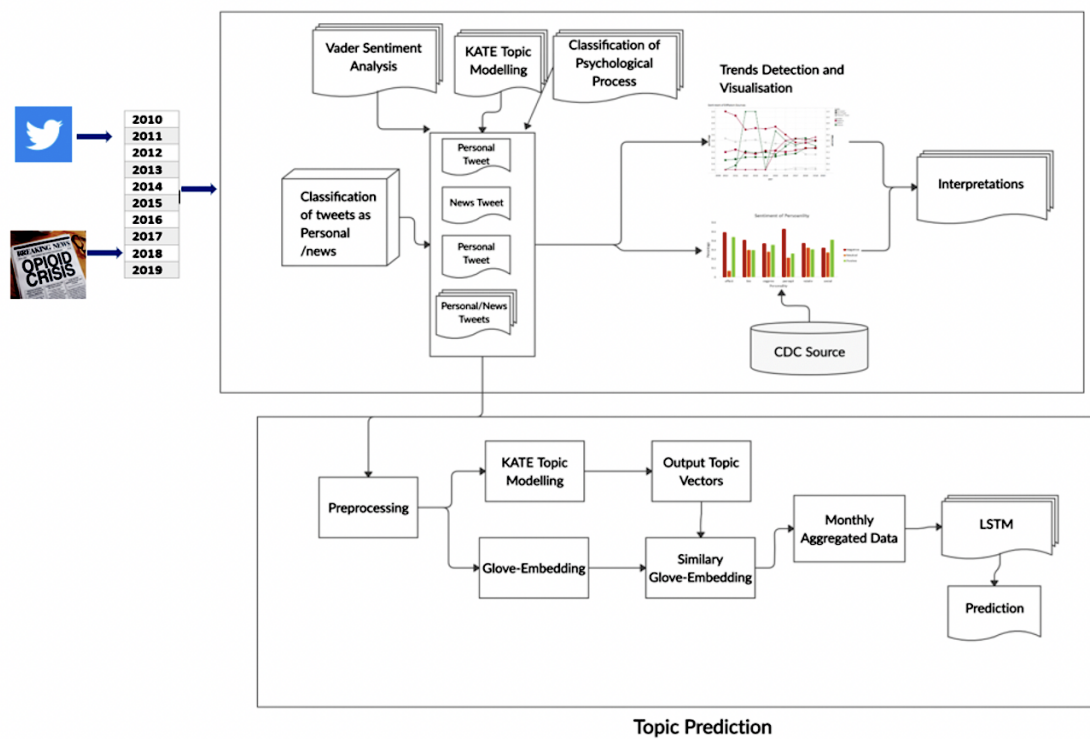


Figure 13: Overall framework of the analysis

Table 21: LIWC Psychological Process

Psychological Process	LIWC2015 Variable included in Distance Calculation
Emotional	affect, posemo, negemo, anx, anger, sad
Social	social, family, friend, female, male
Cognitive	cogproc, insight, cause, discrep, tentat, certain, differ
Perceptual	percept, see, hear, feel
Biological	bio, body, health, sexual, ingest
Motivational	drives, affiliation, achieve, power, reward, risk
Temporal	focuspast, focuspresent, focusfuture
Relational	relative, motion, space, time
Personal	work, leisure, home, money, relig, death
Utterances	informal, swear, assent, nonflu

to avoid overfitting, and g , b and c stand for the activation function, the bias term at the encoder, and the bias term at the decoder, respectively.

Recent autoencoders that perform well on text classification tasks include k-competitive autoencoders, such as K-Spare and KATE. The competition criteria vary from one method to another. For example, K-Sparse aims to enforce sparsity in the hidden layers by keeping the k highest activities during the training phase and αk highest activities during the test phase. K-Sparse uses linear activation functions for the hidden neurons, while the non-linearity in the model derives from the selection of k highest activities. K-Sparse achieved better classification results than denoising autoencoders [44], models trained

with dropout [45], and Restricted Boltzmann Machines when applied for textual data.

KATE (K-competitive Autoencoder for Text) builds on top of K-Sparse for learning meaningful representations by introducing competition among the neurons of hidden layers; KATE's approach is to select k winner neurons composed of $k/2$ largest positive activation and $k/2$ largest absolute negative activations, which then gain the energy of loser neurons.

When applying KATE in this study, we selected the winner neurons from both the strongest and the weakest, the positive and the negative neurons, ensuring more fair competition by providing a second chance to the weakest negative and positive neurons. We also introduced a filtration technique that filtered out similar neurons before entering the competition process. This ensures that the selected winner neurons are distinctive and not redundant.

5.4 Results

The death counts based upon CDC's data **overdoseopioid** in Figure 12 show that the opioid crisis had continued for a few years before it became a major issue of the public discourse. There was a drastic increase in death counts from 2015 to 2016. It coincided with a steeper rise of public discourse on opioids on social media, which peaked in 2017 with the death counts. There was a smaller jump of opioid news stories in nytimes.com in 2016, but the fluctuation of news story counts was less volatile than tweets. The Pearson Coefficient between the numbers of deaths and tweets every year is 0.9, indicating a strong correlation.

Table 22: Topics and Terms

Topics	Terms
Drug Addiction	dare, addition, problem, someone, suffer, acknowledge, time, quit, humili
Drug Treatment	safety, care, opioid, plan, educ, safe, primary, approve, treatment, company
Drug Overdose	death, deter, abuse, advice, gener, increase, drug, overdose, best, tell
Doctor Prescription	drug, doctor, prescript, monitor, overprescribe, Canada, Educ, prescribe, painkiller, target
Obama Plan	Obama, handle, plan, crisis, released
Trump Plan	program, fight, little, maintain, history, critic, trump, hous, already, decent
Drug Makers	fight, crisis, drug makers, opioid limit, limit
Epidemic	epidemic, pain, crisis, address, task, guideline, forc, tune, tackle, combat
Death Rates	patients, overdose, health, years, million, percent, deaths, addiction, products, likely
Marijuana	pharma, drug, epidemic, legal, fight, company, abus, state, marijuana, declare

Ten major topics were identified: Death rates, doctor prescription, drug addiction, drug overdose, drug treatment, drug makers, epidemic, Marijuana, the Obama plan, and the Trump plan. Figure 14 shows that public discourse on Twitter had a largely negative sentiment about death rates, drug overdose, drug prescription, and the Obama plan. In August 2017, the Trump administration declared the country’s opioid crisis a national emergency **achenbach2017trump**. The discussion about the Trump administration’s health policy about the opioid crisis was named “the Trump plan”. Meanwhile, public discourse on the Affordable Care Act (“the Obama Plan”), a major health reform, has continued. The sentiment about drug addiction, epidemic, and the Trump Plan was more neutral. The sentiment about drug treatment was more positive.

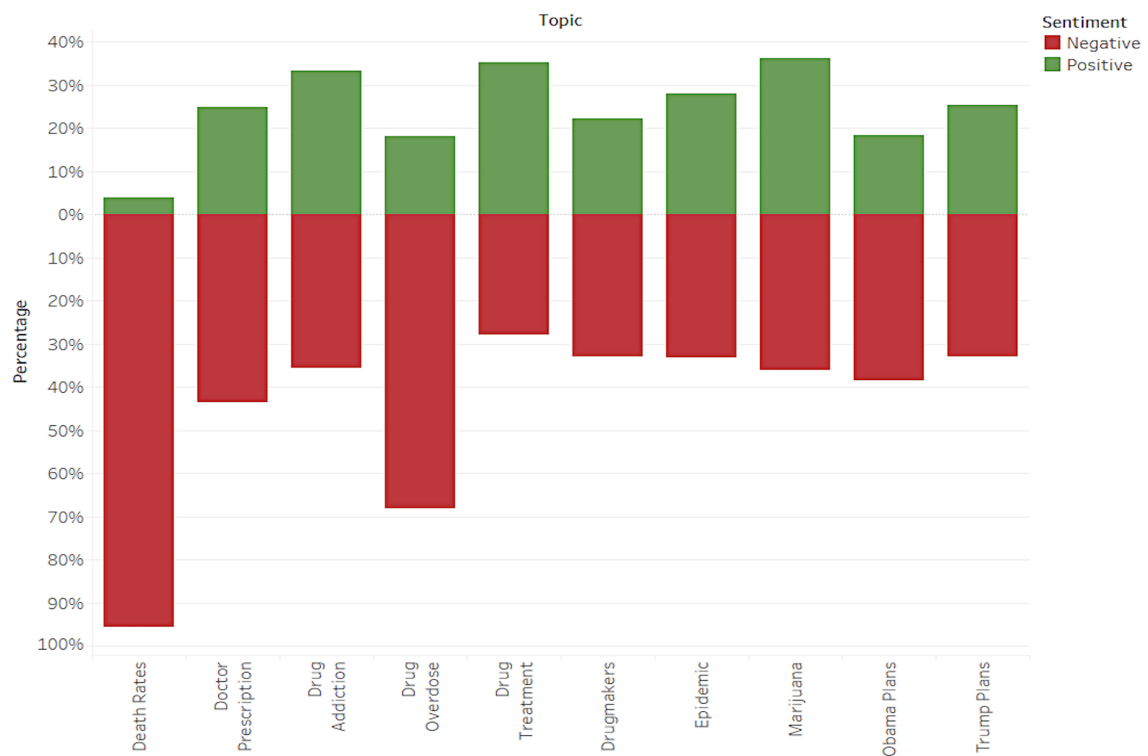


Figure 14: Sentiment of Main Themes of Public Discussion Tweets

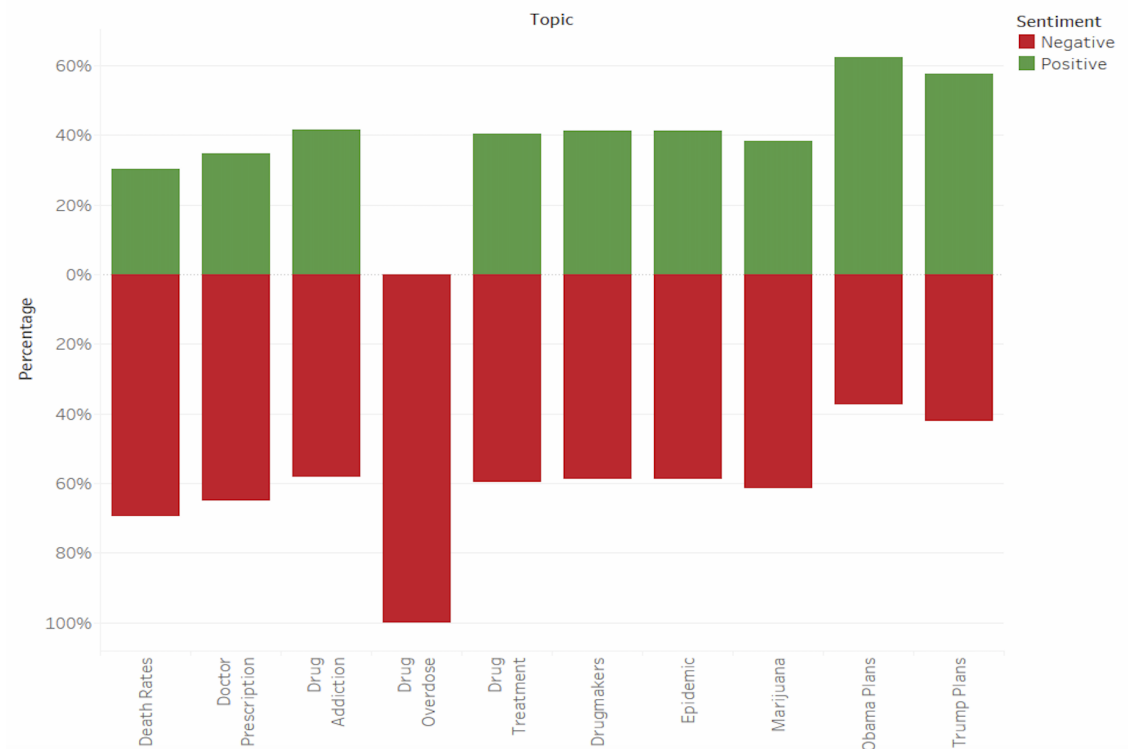


Figure 15: Sentiment of Main Themes of News Story

The sentiment of news stories (see Figure 15) exhibits a more stable pattern than the sentiment of public discourse on Twitter. Reporting health policies (the Obama and the Trump plans) was largely positive, and the reporting of non-policy topics was predominantly negative (drug overdose was entirely negative).

To analyze issues and agendas, time series is essential. However, it takes time for problems and plans to form in the public sphere; they are trends, not events. Thus, time-wise, we used year as the unit of time. The public discourse on Twitter (Figure 16) had a comparatively even distribution of topics, with minimal discussion on health policy issues (the Trump plan and the Obama plan), in comparison to the news agenda (Figure

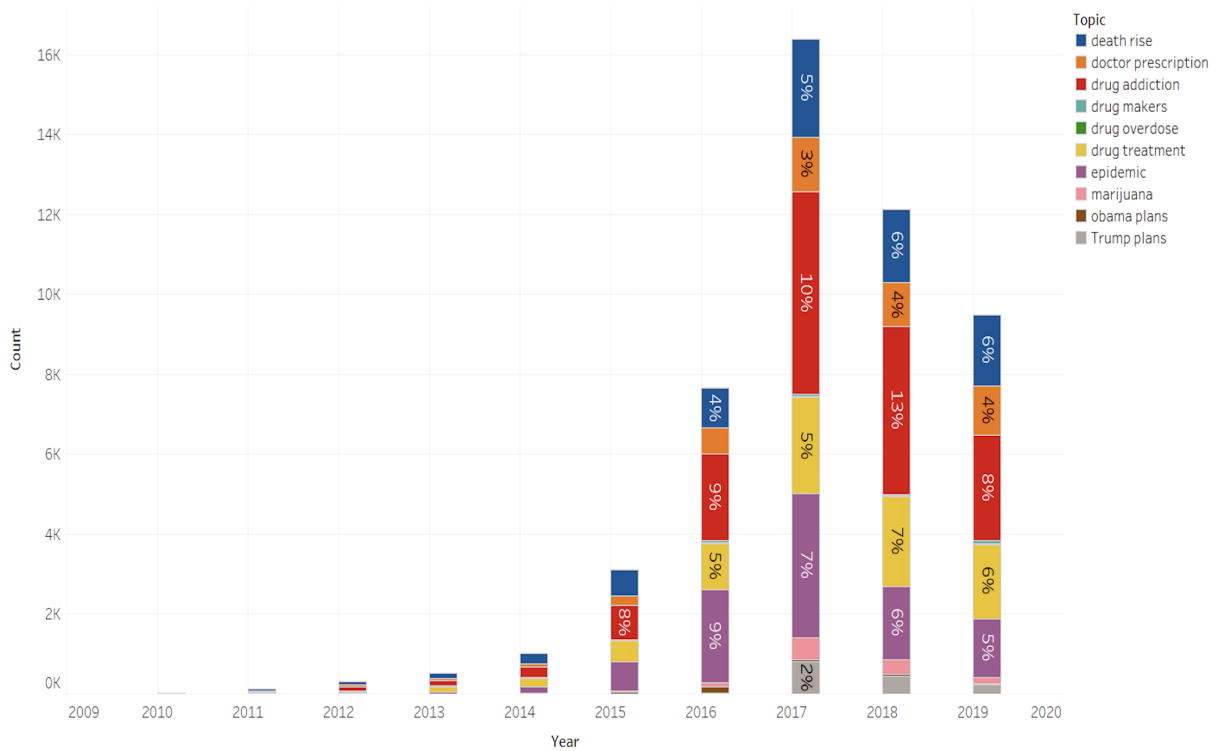


Figure 16: Main Themes of public discussion tweets, 2010-2020

17). Drug addiction had attracted more social media discussion than the other topics. The top five topics were drug addiction, death rate, drug treatment, epidemic, and doctor prescription on Twitter. They were comparatively closer in terms of proportion, and the proportion did not vary significantly over the years.

The news agenda had a significantly more robust focus on policies (the Trump plan and the Obama plan), followed by treatment, death rise, epidemic, and doctor prescriptions. There was a significant increase in policy reporting and a drop in drug treatment reporting from 2016 to 2017.

Personal experience tweets were 2.8% of all tweets. Our analysis conducted a

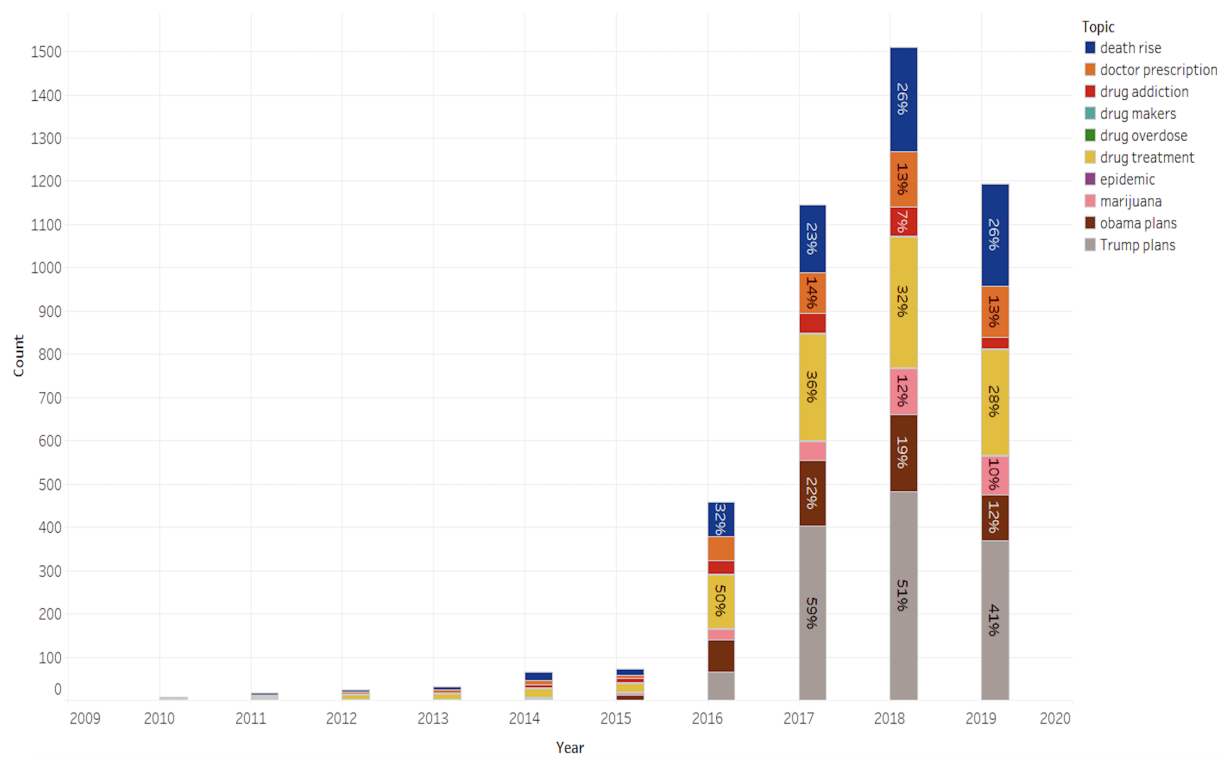


Figure 17: Main Themes of News Story, 2010-2019

close examination of the sentiment of personal experience tweets. To do so, LIWC was using. It is a dictionary that categorizes words to assist analysis of various emotional, cognitive, and structural components present in speeches and writings[90]. Figure 18 shows that there was a significantly high amount of content on drug addiction. When drug addiction was described, biological processes were most frequently mentioned (body, health, sexuality, ingestion), followed by social processes (friends, families, and humans). Figure 21 shows the mention of different types of drugs from 2010 to 2019 in personal experience tweets. Y-axis shows the percentage of tweets. Among all the opioids listed, Morphine and heroin were most mentioned in personal experience tweets.

Drug treatment had a comparatively even distribution of LIWC categories, with the perceptual processes (seeing, hearing, feeling) mentioned the least. Doctor prescription contained more words of the biological processes and the perceptual processes.

In terms of sentiment, the most negative LIWC psychological processes were perceptual and biological processes (see Figure 20). Only the social processes had a similar proportion of positive and negative content. All the other LIWC psychological processes had more negativity than positivity.

5.5 Discussions

From 2014 to 2017, the amount of the public discourse about the opioid crisis on Twitter had a 15-fold increase in volume. From 2015 to 2019, the discourse on Twitter featured themes of drug addiction, death rate, epidemic, drug treatment, and doctor prescription. Comparatively speaking, there was a stronger emphasis on the crisis than on

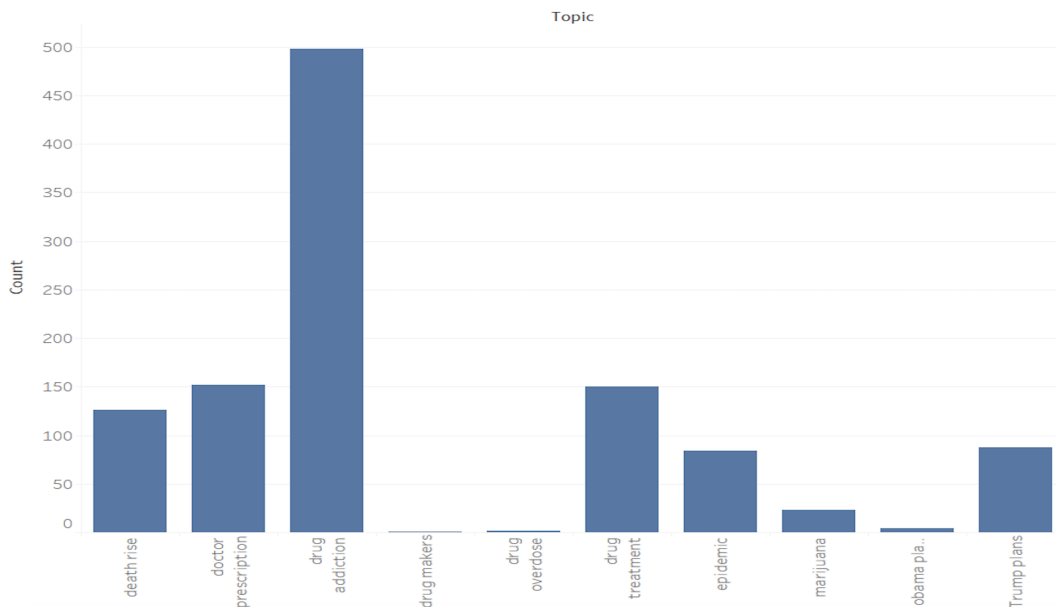


Figure 18: Topics of Personal Experience Tweets

solutions on social media. The top topics were drug addiction, death rate, and epidemic. These top three topics highlighted ordinary users' concerns over the prevalence of the opioid crisis.

The time series analysis suggests that the public discourse on social media identified the opioid problem earlier than the news media. The volume of public discussion tweets peaked in the same year (2017) when the death counts peaked. The discussion of drug addiction had the highest volume in 2017, 2018, and 2019 on Twitter. In contrast, drug addiction was not the top issue of news reporting on nytimes.com over the years, and the peak of opioid news stories came a year (2018) after the death counts peaked. This implies social media platforms are quicker than the news media to raise concerns over public health problems.

Meanwhile, our findings suggest that collective meaning construction took place

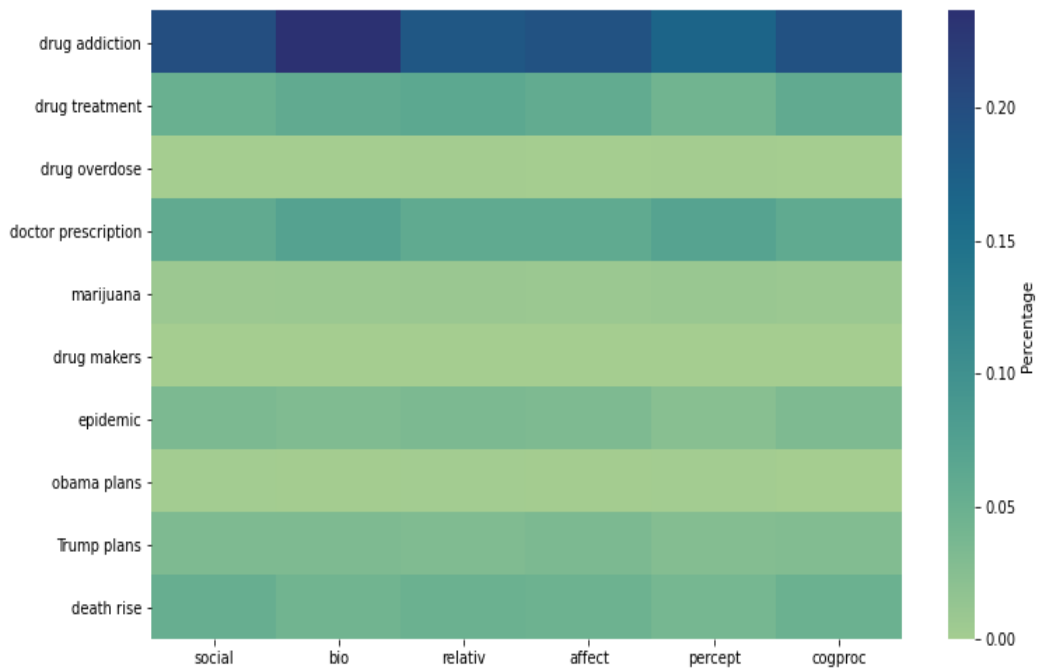


Figure 19: LIWC Distribution in Personal Experience Tweets

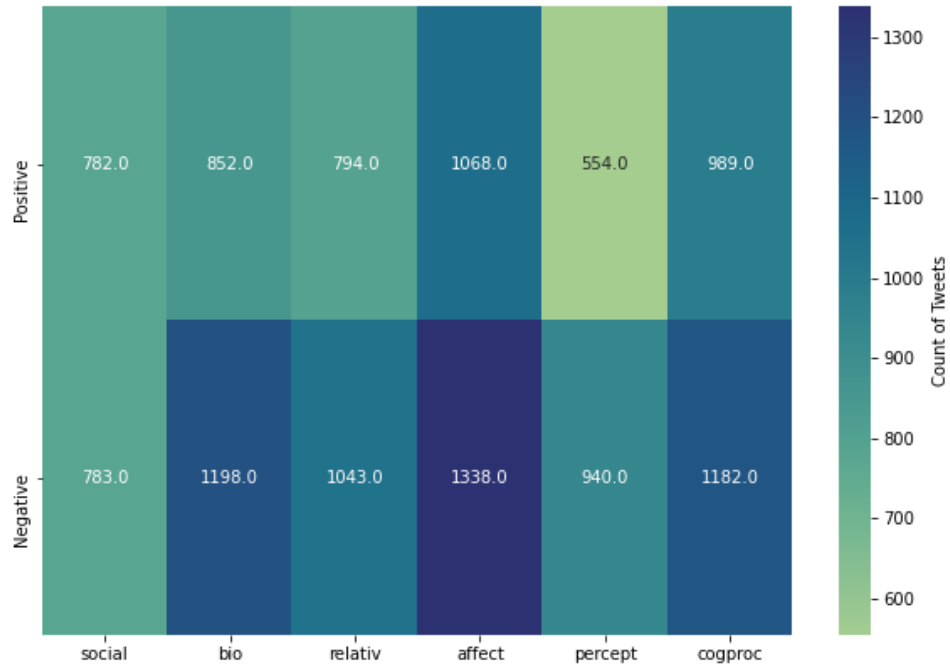


Figure 20: LIWC Sentiment of Personal Experience Tweets

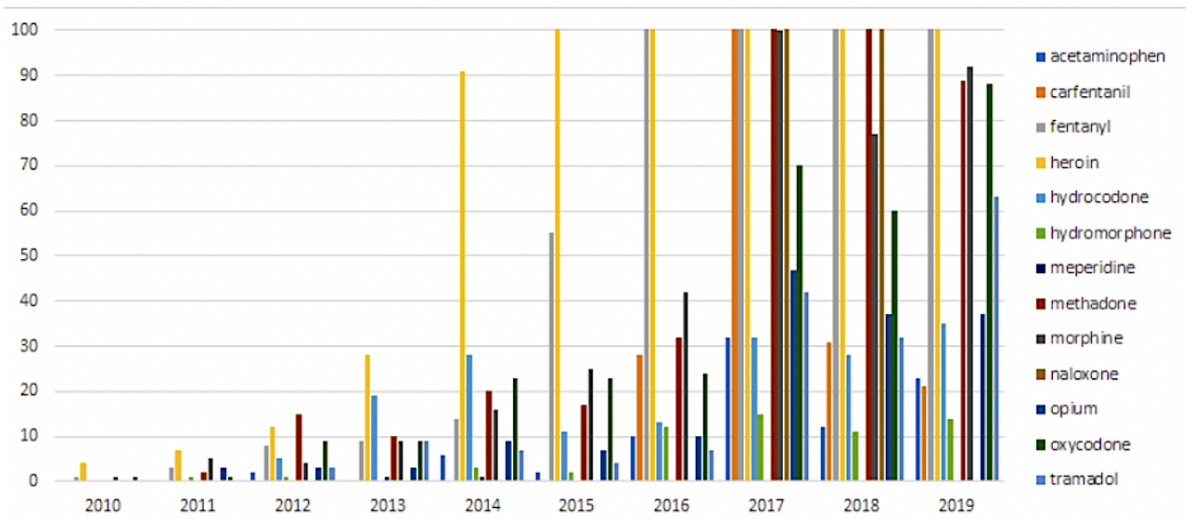


Figure 21: Mention of Drugs in Personal Experience Tweets

on social media. First, Twitter’s public discourse featured more diverse topic interests in comparison to news stories. The proportions of topics were distributed more evenly than that of news stories. Second, there was minimal discussion about health policy on Twitter, while it dominated the opioid news since 2017. As discussed before, the social media discourse highlighted the prevalence of the addiction problem. This contrast suggests that public discourse on social media was geared more toward identifying the problem with some discussion about individual-level solutions. Its agenda and frames were an aggregation of individual inputs, and thus was less deliberate than the news media. The news media apparently emphasized macro-level solutions, such as health policies.

In line with previous studies, evidence from this study strongly suggests that public discourses on social media have their own “agendas.” Social media is where the public collectively defines what constitutes the meaning of the “opioid crisis.” The public discourse taking place on social media represented diverse perspectives and was more spontaneous to the pain and the damages caused by the opioid crisis. The collective sentiment (predominantly negative) suggests a strong desire for change. This tendency of being reactive may explain why long-term solutions such as policy solutions were not the central theme. In contrast, news media’s agenda-setting and framing seem to be a more deliberate process. It focused on, in this specific case, the solutions rather than identifying the problem. For example, in 2016, 50% of the news stories were about drug treatment, which is a medical solution to the crisis. The focus switched to the Trump Plan in 2017, which is a policy solution. These highly focused topics and the apparent shift of topics from 2016 to 2017 are vital signs that journalists practiced reporting more uniformly to consciously

push medical and policy solutions. The policy solution was framed more positively than negatively, suggesting the news media's encouragement to policy-level solutions.

The last issue to discuss is the nature and characteristics of “personal experience tweets.” Existing studies that track incidences of diseases used the first-person pronoun. Instead, this study used a clue-based approach to exclude subjectivity and profanity, and draw the boundaries for public discourse. Given that the purpose of the study is to examine public discourse, it is more appropriate to focus on the content rather than grammatical markers. To explore the part of the data that did not belong to the “public discourse”, we conducted a more subtle sentiment analysis on the “personal experience tweets.” First, these tweets heavily focused on drug addiction. The LIWC sentiment analysis implies that personal experience tweets about drug addiction included a lot of content about the physical pain associated with drug abuse (negative sentiment of the biology category of LIWC). There were also signs of social support from families and friends, which can explain a more neutral sentiment of the social process of LIWC. The evidence suggests personal experience tweets include patients and their family and friends' writing about the direct impact of opioid abuse on individuals.

5.6 Conclusions

Given the patterns of topics, sentiment, and time series, public discourses of the opioid crisis on social media documented responses directly from ordinary Twitter users in a timely manner. To the ordinary users, the opioid crisis has caused so many deaths and tragedies; it is a serious health issue that needs to be addressed and solved. In general, the

social media agenda can react to the crisis quicker than news reporting, and feature more diverse perspectives. Thus, analyzing social media public discourses can help raise early warning signs for public concerns over health issues.

CHAPTER 6

EARLY TEMPORAL CHARACTERISTICS OF ELDERLY PATIENT COGNITIVE IMPAIRMENT IN ELECTRONIC HEALTH RECORDS

6.1 Introduction

Medical advancements have resulted in an increase of 30 years in the average longevity of the people since the turn of the twentieth century. In 2012, the United States had 40.7 million persons aged 65 and older (13.2 percent of the total population), with 38.7 percent reporting having one or more impairments [135]. The aging population has also led to an increase in persons living with CI, more than 17 million people in the United States [136], causing patients, their families, and society an annual estimate of \$18 billion in lost income and direct cost of care [137]. We define cognitive impairment (CI) in this section as either moderate cognitive impairment (MCI) or dementia. According to Alzheimer's Disease International, dementias impacted 46.8 million people globally in 2015. They projected that the number would nearly triple to 131.5 million by 2050 [138]. Regarding this, the subject of MCI is paramount as it is a transitional zone between normal life in older ages and dementia. One study indicated that clinicians were not aware of CI in more than 40% of their patients [139]. Failure to diagnose cognitive symptoms would delay the implementation of effective treatment strategies for underlying illnesses and coexisting disorders, and may result in safety concerns for patients and others [140], [141]. In many cases, the CI problem will worsen over time [141]–[143]. Thus, early

diagnosis of CI can be of utmost importance and may reduce the large burden later on the medical and social care.

The impact of CI on ADL has been used as a criterion to differentiate MCI and dementia [144]. ADL is often divided into basic ADL (b-ADL), which includes activities such as personal hygiene, clothing, feeding and toileting [145] and instrumental ADL (i-ADL), which is commonly referred to as independent living abilities such as household activities, handling money, shopping, and transportation [146]–[148]. The i-ADL has a higher demand for cognitive function than the b-ADL and is important for living an independent life in society [149]. ADL is highly dependent on cognitive function and behavior [150]. Therefore, there should be assessments that are capable of detecting changes in ADL as soon as changes in cognition and behavior are detected [150]. We began our investigation by examining the fundamental statistics of an EHR corpus important to CI diagnosis. Temporal patterns in ADL were examined between CI and CU patients in senior patients (age 65 or older) mined from EHRs prior to developing CI. We analyzed both structured (patients' current visit information) and unstructured data (clinical notes). Additionally, we used machine learning approaches (three topic modeling methods) to extract relevant semantics (i.e., subjects and phrases) from clinical free text in order to investigate their possible relationship with future CI development.

Different studies have used machine learning algorithms to differentiate between cognitively normal and MCI individuals [151], [152], to predict conversion from MCI to Alzheimer's disease (AD) [153], and to predict the time to this conversion [151]. Researchers [154] devised a two-layer model, with the first layer serving as a screening tool

for determining if a group is normal or aberrant. The second layer is a detailed evaluation to determine whether the patient has MCI or dementia. They compared the results using a variety of machine learning techniques. High performance was demonstrated using support vector machines, multi-layer perceptrons, and logistic regression. Additionally, conversion from MCI to AD has been explored using a deep learning model in conjunction with MRI, neuropsychological, and demographic data.[155]. In another study [156], they tried to predict MCI from spontaneous spoken utterances. Classifying cognitive profiles using machine learning with fMRI data as an addition to cognitive data were explored [157]. In their work fMRI data are only used to train the classifier and classification of new data is solely based on cognitive data. Another research [158] focuses on the early detection of Alzheimer's disease using deep learning and a sparse autoencoder. They identified regions of the brain that are vulnerable to AD progression using neuroimages derived from the neuroimaging initiative database. These prior research attempted to extend their findings by include fMRI data in their models. Although it may have a beneficial effect on the outcome, not all patients have fMRI data, and so the application may be limited in scope when compared to the application employing standard EHR data in the health care community. Additionally, they did not attempt to find novel risk variables linked with CI patients in EHRs, instead relying on already identified medical illnesses and fMRI data to predict CI.

There are studies focused on predicting progression from MCI to dementia using neuropsychological data. Researches [159] evaluated the results of neuropsychological tests in order to determine their utility for predicting dementia using a machine learning

system. They employed a feature selection ensemble technique to identify the neuropsychological test variables that were predictive of acquiring AD dementia. Additionally, the neuropsychological test used to estimate time conversion from MCI was explored in [160]. In this study, MCI patients were classified according to whether they progressed to dementia (converter MCI) or stayed stable (stable MCI) during a specific time period. After that, a prognostic model was built to forecast the conversion time up to five years prior to acquiring dementia.

In contrast to earlier research, we employed a machine learning technique (i.e., topic modeling) to evaluate themes and phrases in EHR free text that may be useful for early diagnosis of CI. A few research have focused on the early detection of CI [161]; however, these studies employed traditional methods such as i-ADL and b-ADL assessments rather than machine learning algorithms using EHR free text.

6.2 Method

We analyzed the fundamental EHR corpus statistics (i.e., distributions of event types and practice settings associated with the first CI diagnosis, and the quantity of clinical notes between CI and CU patients). Temporal trends in patient ADL were compared and topics in clinical free text were analyzed over time between physician-diagnosed CI and CU patient groups using three machine learning models.

6.2.1 Data

The study cohort was drawn from patients 65 years of age or older at the time of enrollment in the Mayo Clinic Biobank (n = 22,772), where we identified physician-diagnosed CI (n = 1,435; 55% male) and CU (n = 1,435) patients who were age (+/ 1 year) and sex matched. Physician-diagnosed CI patients were identified using the diagnostic section of clinical notes (i.e., dementia, cognitive impairment, cognitive deficit, cognitive decline, moderate cognitive impairment) [162].

6.2.2 Corpus Analysis

The fundamental EHR corpus statistics related to CI diagnosis were evaluated (i.e., the distributions of event types and practice locations associated with the first CI diagnosis), as well as the quantity of clinical notes over time for CI and CU patients.

6.2.3 Analysis of Activity of Daily Living

Two sources were used to compile the ADL: 1) current visit information, which patients supply and update every six months during their visits to the Mayo Clinic; and 2) specific areas of clinical notes (i.e., instructions for continuing care, ongoing care orders, system review). The present visit information contains structured questions to measure patients' capacity to do ADLs (binary evaluation measuring the difficulty of ADLs: yes or no). To extract ADL-related concepts, the clinical notes were processed using the Med-TaggerIE module in MedTagger [163], [164], which is an open-source pipeline developed by Mayo Clinic for pattern-based information extraction with assertion detection (i.e.,

negated, possible, hypothetical, associated with a patient). Through the MedTaggerIE implementation, these notions were automatically transferred to the respective preset ADL categories (i.e., rule-based normalization process). We included just non-negated ADL notions.

Once we obtained ADL concepts, they were mapped to items in Katz's index (b-ADL) [145] and [146] scale (i-ADL), which are the most commonly used tools for assessing ADL. The items of ADL used in this study for each ADL category are: 1) b-ADL: bathing, dressing, transferring, toileting, and feeding; 2) i-ADL: using transportation, shopping, preparing food, housekeeping, responsibility for own medications, and handling financing. These items can be mapped to the International Classification of Functioning, Disability, and Health (ICF), allowing for broad information exchange. The temporal trends of b-ADL and i-ADL between CI and CU patients were compared in every 6 months for 5 years before the first physician-diagnosed CI and the latest visit for CI and CU patients, respectively.

6.2.4 Analysis of Topics in Clinical Notes

The themes in clinical notes were examined in two ways: 1) how topic terms have evolved over the last five years in CI patients (experiment 1), and 2) how topic terms differ between CI and CU patients throughout the five years preceding the formation of CI (experiment 2). (experiment 2). This step-by-step timeline enables us to watch how the issues evolve over time, motivated by the expert opinion that persons over the age of 65 should see their doctors every six months to evaluate if their symptoms are

remaining stable, improving, or worsening [150]. We examined 1) entire clinical notes, 2) individual sections (i.e., history of current illness, diagnosis, current medication), and 3) the set of sections most likely to contain medical concepts of interest (i.e., chief complaint, history of current illness, system review, prior medical history, physical examination, impression/report/plan, and diagnosis).

After eliminating stop words and stemming, we use the most common 2,000 words as the vocabulary for preprocessing the topic models. As follows, we used three distinct machine learning models: two traditional topic modeling techniques (LDA and TKM) and one deep learning technique (KATE). The topic count was established using the self-regulatory capacity built in a TKM model.

6.2.4.1 Latent Dirichlet allocation (LDA)

It is a probabilistic generative model in which the document is considered as a collection of diverse topics, with each topic representing a distribution of the word. We limited ourselves to twenty subjects and a ten-word distribution within each topic. Other hyperparameters were set in accordance with the code in [33].

6.2.4.2 Topic keyword model (TKM)

This method addresses the shortcoming of LDA approach (i.e., ignoring the order of words). In TKM, each word in each topic aims to show how common the word is within the topic and how common it is between other topics [42]. The other advantage of this method is that redundant topics will be removed automatically. We used the hyper

Table 23: Average number of clinical notes for CI and CU patients (SD in parenthesis)

Year	CI patient	CU patient
1	30.7 (41.5)	21.8 (34.0)
2	22.4 (26.4)	17.2 (24.0)
3	21.4 (25.7)	17.0 (23.5)
4	21.1 (25.7)	16.1 (20.9)
5	18.7 (21.3)	15.4 (18.7)

parameters as explained in the paper in [42].

6.2.4.3 K competitive autoencoder (KATE)

An autoencoder is a type of neural network that is capable of learning data representations autonomously by synthesizing its input at the output level. Numerous autoencoder variations have been proposed, mostly for picture data. However, KATE was developed to address the shortcomings of classic autoencoders, which are incompatible with textual data citechen2017kate. The number of topics in this experiment was limited to 20, with each topic receiving a distribution of ten words. Other parameters for deep learning were established as mentioned in the original study [21].

6.3 Results

We began by examining the cohort's basic EHR corpus data. Then, we evaluated temporal changes in 1) b-ADL and i-ADL, as well as 2) individual ADL, in CI and CU patients prior to the development of CI. Three topic modeling methods were used to analyze and compare the terms and topics extracted from clinical notes between the two patient groups over time, both qualitatively and quantitatively, in order to gain a better

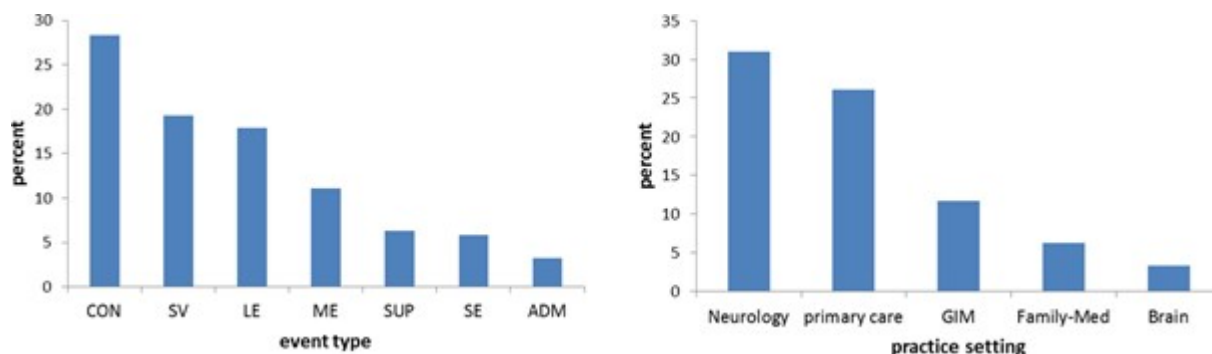


Figure 22: Distribution of the first CI diagnosis (CON: consult, SV: subsequent visit, LE: limited exam, ME: multi-system evaluation, SUP: supervisory, SE: specialty evaluation, ADM: admission; GIM: general internal medicine)

understanding of the patient medical conditions that may contribute more to CI development.

6.3.1 Corpus Statistics

22 shows major event types (i.e., note types) and practice settings along with their occurrences in which a physician first diagnosed CI. The consultation was the most dominant event to diagnose CI (28%), followed by subsequent visit (19%), limited exam (18%), multi-system evaluation (11%), and supervisory (6%), which cover more than 80% of total events of CI diagnosis. For practice setting, neurology (31%) was the most dominant, followed by primary care (26%), general internal medicine (12%), family medicine (6%), and brain (3%). Table 1 contains the statistics of clinical notes for the past 5 years of CI and CU patients before they develop CI and the latest visit date, respectively. As can be seen, CI patients consistently showed higher reading of clinical notes than CU patients and the difference was most significant in the first year before CI diagnosis.

6.3.2 ADL Distribution

Figure 23 depicts the temporal distributions of CI and CU patients' worsened b-ADL and i-ADL in three age groups (65-74, 75-84, and 85+). In general, CI patients had worse b-ADL and i-ADL (i.e., a larger ratio of degraded ADL) than CU patients across all age groups, and this tendency is more pronounced when the CI patients are close to being diagnosed with CI. The deterioration of b-ADL and i-ADL is similar between the age ranges of 65-74 and 75-84 for both CI and CU patients. Interestingly, whereas the b-ADL of CU patients was often worse than the i-ADL, the reverse was true for CI patients—i.e., CI patients' i-ADL deteriorated with time, particularly close to 1.5 to 1 year(s) before the physician identified CI.

Additionally, we compared the ADL trajectories of the total patient group to those of CI and CU patients. Across all ADL categories, CI patients had more worsened ADL than CU patients with time. Transfer (17 percent for CI patients and 14 percent for CU patients) and housekeeping (14 percent for CI patients and 10% for CU patients) were the most deteriorated ADLs six months before. The difference between the two groups is little when it comes to cleaning and transfer, but significant when it comes to bathing and responsibility for one's own medicine 24.

6.3.3 Topic Modeling

6.3.3.1 Qualitative Analysis

We compared hidden themes in CI patients prior to the onset of CI using topic words extracted by three different models (i.e., LDA, TKM, and KATE) from clinical

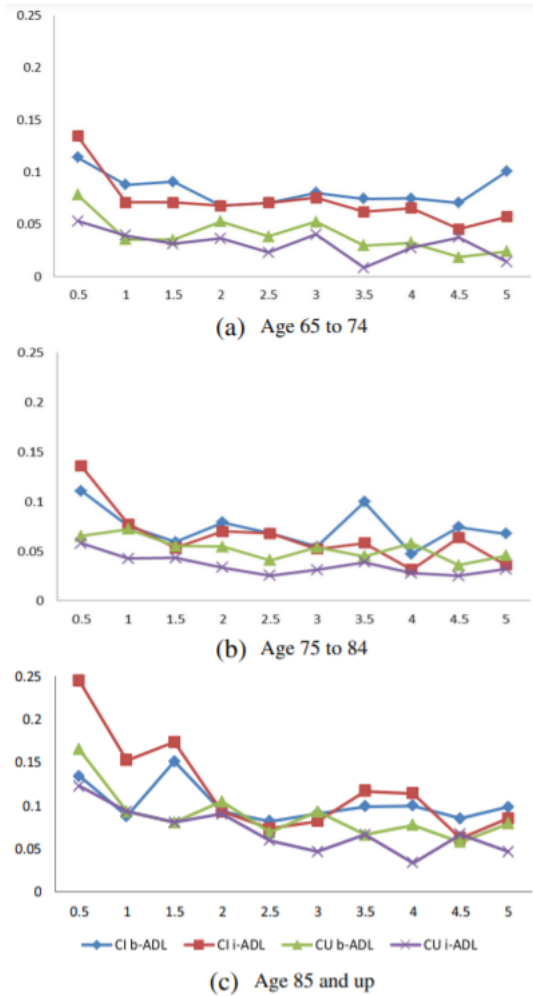


Figure 23: Distribution of b-ADL and i-ADL for CI and CU patient groups (x-axis is year(s) before the 1st physician-diagnosed CI for CI patients and the latest visit for CU patients; y-axis is a ratio of patients who have a deteriorated ADL)

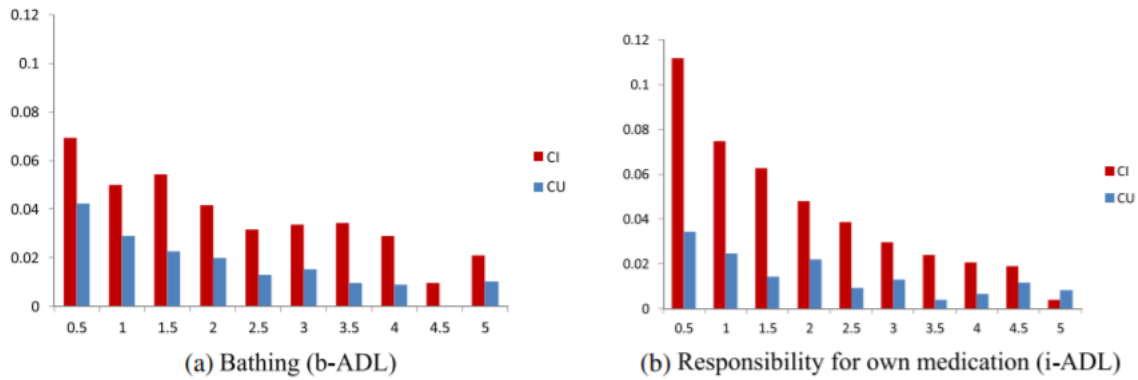


Figure 24: ADL distributions for CU and CI patient groups (x-axis is year(s) before the 1st physician-diagnosed CI for CI patients and the latest clinical visit for CU patients; y-axis is a ratio of patients who have a deteriorated ADL)

Table 24: Topic words by TKM (6 months before CI diagnosis)

Section	Word distribution for topics
All	pain symptom scalepati feel hpi fatigu numer loss vomit rate worst appetit statusth climb headach
Set of sections	sleep apnea cpap obstruct sleepi oximetri interfac daytime polysomnographi snore
History of present illness	glucos sugar metformin pressur blood insulin vitamin diabet cholesterol losartan interact hydrochlorothiazide
Medication	sugar glucos metformin decitabin blood pseudogout copeman diabet losartan fast insulin lantu read station glipizide
Diagnosis	lesion carcinoma cell dermatolog melanoma ulcer cancer squamou surgeri concern examin nonmelanoma

Table 25: Topic words by KATE (6 months before CI diagnosis)

Section	Word distribution for topics
All	mouth hpi releas capsul sleep apnea gi nasal prescript obstruct
Set of sections	medic prescript sleep hypertens obstruct concern diabet apnea hyperlipidemia acut
History of present illness	chronic diseas atrial hypertens failur fibril acut heart coronari back
Medication	drop zocor aspirin ophthalm day tablet low atenolol apr
Diagnosis	histori sleep apnea obstruct hypertens disord hyperlipidemia neuropathi bilater depress

Table 26: Topic words by LDA (6 months before CI diagnosis)

Section	Word distribution for topics
All	normal distress clear alert bilater soft edema sound orient tender
Set of sections	diseas histori hypertens chronic statu arteri atrial hyperlipidemia coronari diabet
History of present illness	urinari urin incontin bladder infect tract symptom deni histori urgenc
Medication	carbidopa levodopa hs benjamin start vitamin garlic bid knutson hydrochlorothiazide
Diagnosis	memori hypertens concern hyperlipidemia health mainten chronic hypothyroid complaint elev

notes. This method may help identify probable patient medical issues that contribute to CI. Table 24, 25, 26 incorporate topic phrases produced by various topic models in various sections of clinical notes during the six months before the physician’s diagnosis of CI. The strong type indicates the linked terms in a specific issue that are important to CI. The tables include stemmed words. For each part, we presented a typical cluster of the themes. As seen in the tables, the themes are distinct from one another, resulting in a meaningful representation of the text data. For example, ??, all sections show some symptoms related to ”fatigue”, which may be the potential risk of dysfunction [165]; the topic in set of sections is relevant to ”sleep issue” that could be observed in the individuals suffering from cognitive disorder [165], [166]. The topic words in the history of present illness section, we can observe glucose, diabetes, insulin, and hydrochlorothiazide, which are related to diabetes disease considered as a potential risk factor of cognitive decline [165]. For the topic in the medication section, we observed medications to control high blood sugar [165]. The topic in the diagnosis section includes the terms related to cancer [167]–[170]. Table 25, set of section and history of present illness include hyperlipidemia

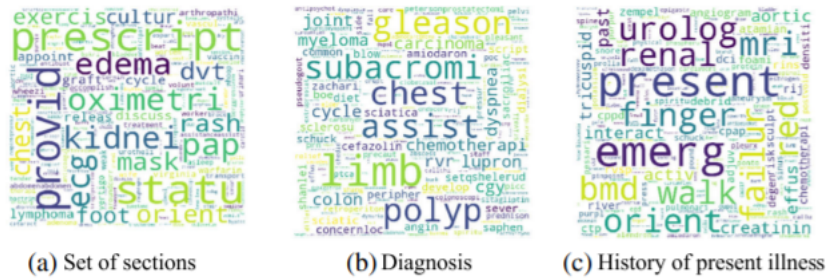


Figure 25: Topic terms for CI patients - TKM (Experiment 1)

that can be considered as a risk factor of CI [171], coronary artery disease and hypertension, which are relevant to cognitive decline [171], [172]. In Table 26, LDA result in similar outcomes as TKM and KATE is shown. Words like edema, distress, memory, hypertension, coronary, urinary and hyperlipidemia as the potential risk factor of cognitive dysfunction was discussed [172]–[175]. Carcinoma, melanoma, cancer, and squamous in the last row are the terms related to cancer **magnuson2016ajean2010management**, [169], [170].

6.3.4 Quantitative Analysis

We assessed how the subject words learnt by the topic models have altered in CI patients as they approached physician-diagnosed CI over the last five years (experiment 1), and how they have remained unique between CI and CU patients over the same time period (experiment 2). (experiment 2). We analyzed the cumulative phrase frequency in topic terms over time. For the first technique (experiment 1), the variations in topic phrase frequencies between two consecutive years preceding the CI diagnosis were computed (beginning one year previous to the CI diagnosis), and then repeated for each year during a five-year period. Each year, we employed 400 topic phrases. This may help

us find prospective subject phrases connected with CI development, since we may detect more frequent occurrences of CI-related terms as the CI diagnosis date approaches. We employed the same aggregated term-frequency difference technique (experiment 2) for the second approach (experiment 2), but for the whole 5-year period. This method, the phrases that are often used by CI and CU patients may be separated, and the remaining terms are likely to be those connected with CI. We utilized the whole five years because we did not find any relevance when comparing year by year.

Figure 34 shows the high-level concept of our approach using aggregated term differences. The result of these approaches is visualized in Figure 25, 26, 27, 28, 29 and 30. The larger words denote that they appear more frequently in the result of topic modeling on clinical notes compared to the previous year (experiment 1), or in the whole 5 years (experiment 2) (the corresponding individual raw data in Figures 25, 26, 27, 28, 29 and 30 are located in Tables in Appendix). The results were compared with the recent publication to verify whether this approach generates meaningful outcomes relevant to CI. A disease, "lymphoma" was seen in multiple results (Figures. 25a, 26b, c, 27a, b, c, 28b, 29b, c and 30a, c), which appeared in Hodgkin lymphoma patients complaining about cognitive deterioration and fatigue [176]. A researcher found that cognitive decline was more severe and frequent in Hodgkin lymphoma patients compared to the healthy population [176]. Based on recent study patients with "nocturnal hypoxia" had poor memory retention compared with healthy individuals [177]. Indeed, "oximetry" (Fig. 25a) is a device able to measure the oxygen saturated in the blood in hypoximia patients. In another study [178], a researcher demonstrated that "global cerebral edema" is a vital risk

factor for cognitive dysfunction which we see more frequently in Figs. 25a, 26a, and 30b,c. Researchers studied the association between cancer and cognitive decline in older ages [172]–[175]. They concluded that cancer therapy could negatively impact cognition in some patients. Regarding to this, the word "metastasi", "squamous", "chemotherapy", "oxaliplatin", and "carcinoma" can be seen in Figs. 25c, 27b, c, 28a, b, 29b, and 30a, c. It has been explored that "tinnitus patients" are more at risk of the cognitive deficit as shown in Fig. 25b, c [179]. The word "bevacizumab" in Fig. 28a is a cancer medicine that interferes with the growth of a cancer cell in the body. Indeed, it is used to treat certain types of brain cancer or kidney cancer. The relation between urinary disease and CI has been investigated in several studies (Figs. 26c and 28b) [174], [175]. The words like "depression", "confusion", "memory", and "pressure", which has been already known as the sign of CI can be seen in the Figs. 26b, 27a, b, c, 29b, c, and 30b, c. A couple of the studies explored the relationship between CI in late life and hyperlipidemia, hypertension, and coronary (Figs. 26a, 29a, 28c, and 30c). Heavy snoring and sleep apnea in Figs. 26a, b, c and 28a have been investigated largely by researchers which shows a strong link to earlier cognitive decline [165]. An apnea/hypopnea index is an index, which is usually used to indicate the severity of sleep apnea in patients, is another extracted topic repeated 8 times more in the CI population compared with CU. CPAP is used to treat sleep-related breathing disorders including sleep apnea (Fig. 28c). Diabetes diseases have been identified as a potential risk of cognitive dysfunction [166] and regarding that topic diabetes, glucose, and sugar [172], [173], [180] can be seen at Figs. 26c, 29c, and 30a. In [181], researchers showed that memory impairment has a particular association with the presence

6.4 Discussion

It is critical to recognize early indicators of CI so that physicians may plan appropriately and take necessary action, alleviating possible expense and suffering. The purpose of this study was to examine basic EHR corpus statistics relevant to CI patients and to compare the temporal trends of patient ADL over time and clinical note topics between CI and CU patient groups in order to characterize and better understand the medical conditions of elderly patients prior to the development of CI. The consultation was the most significant event type, and neurology was the most prevalent practice setting in which doctors initially diagnosed CI. The continuously greater quantity of clinical notes for CI patients compared to CU patients implies that CI patients visit hospitals or clinics more frequently than CU patients. Individual ADL and ADL groups (i.e., b-ADL and i-ADL) were investigated throughout time in the five years preceding the first physician-diagnosed CI and the most recent visit for CU patients, respectively. The trajectories of ADL deterioration grew steeper in CI patients than in CU patients roughly 1 to 1.5 year(s) before to the physician diagnosis of CI. Notably, i-ADL degradation was greater than b-ADL deterioration in CI patients over this period, which was not the case in CU patients. Given the large delay in CI diagnosis and the missed chance for proper planning in current practice, this observation may help encourage early detection of CI [138]. Bathing (b-ADL) and self-medication (i-ADL) trajectories deteriorated much more rapidly in CI patients than in CU patients across time. These metrics may also serve as a proxy symptom to aid in the early detection of CI. The findings of this study indicate that topic modeling might be beneficial for identifying relevant and hidden subjects and phrases in clinical notes.

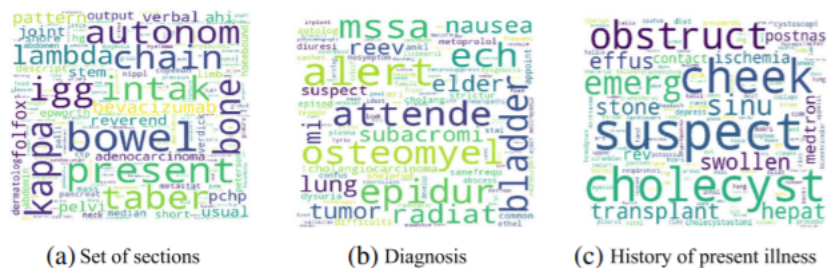


Figure 28: Topic terms in the TKM model (Experiment 2)

As described in the qualitative and quantitative analyses, the outcome was encouraging. We discovered that the majority of the terms in the issue were correlated and accurately conveyed the underlying semantics. The program was able to extract terms related with CI, such as hypertension, depression, and memory, which are all possible indicators of CI. Additionally, we identified other possible risk factors for CI based on recent publications.

Tables 5, 6, 7 contain sample raw data in experiment 1, corresponding to Figs. 5, 6 and 7. It illustrates how many times of a given term appeared more in clinical notes of CI patients than the previous year. Tables 8, 9, 10 contains sample raw data in experiment 2, corresponding to Figs. 8, 9 and 10. It illustrates how many times a given term appeared more in clinical notes of CI patients than CU patients for the past 5 years. In the following tables, "Rate of increase" denotes number of times for a given term appeared more for the first year before CI diagnosis compared to the previous year (Tables 5, 6, 7) or for the past 5 years (Tables 8, 9, 10).

In general, current models TKM and KATE performed better than LDA in capturing the semantically meaningful representation of the data. Additionally, the KATE model yielded more terms associated with CI, including those associated with memory, depression, hypertension, dizziness, and confusion, than the TKM model. On the basis

which does not discriminate the severity of CI, rather than a comprehensive evaluation or test owing to their unavailability. However, our study is still relevant since it examines the use of EHR documentation to encourage early identification of CI, which is critical given doctors' considerable delay in diagnosing CI in current health care practice. Another possible drawback would be an unequal distribution of clinical notes for certain disorders (e.g., cancer patients are seen more than others and have more clinical notes). This may have an effect on the outcome of topic modeling; nonetheless, we evaluated a diverse variety of themes and proved a high degree of possible application.

6.5 Conclusion

There are significant variations in the temporal patterns of b-ADL and i-ADL between CI and CU patients roughly 1 to 1.5 years before physician-diagnosed CI*—i.e.*, the steeper slope of total ADL deterioration and poorer i-ADL than b-ADL in CI patients during this period. Individual ADL (bathing and self-medication) trajectories were found to be strongly linked with the development of CI. Over time, the themes and phrases extracted from clinical free text using topic modeling approaches have

the ability to demonstrate how CI patients' illnesses progress over time and to uncover previously undiagnosed conditions as they near CI diagnosis. These findings may contribute to the early diagnosis of CI and consequently to the expedited treatment of associated illnesses and concomitant disorders. In the future, we intend to use neuroimaging and assessment data to establish a more precise classification of cognitive function and to develop a prediction model based on our observations to identify patients at risk of

developing different stages of CI and to identify associated longitudinal risk factors.

CHAPTER 7

MINING NEWS MEDIA FOR UNDERSTANDING PUBLIC HEALTH RECORDS

7.1 Introduction

Identifying sentiments and focuses of news media toward public health concerns is an emerging research topic of interest. News media play a substantial role in raising public awareness, framing public opinions, and affecting policy formulation and adoption of popular issues [186]–[188]. In the area of healthcare, news media use multiple channels to communicate evidence-based research findings to individuals and healthcare professionals and accelerate the translation of these research findings in healthcare to public health practice. For example, the news media have a drastic impact on changing the public's perceptions, attitudes, and behaviors toward smoking, alcohol-impaired driving, and healthcare service utilization [189]. News media have a tendency to use language to influence the public's opinions, behaviors, and perceptions related to specific health issues. For instance, antismoking articles emphasized the health risks of smoking with negative sentiment (e.g., fatal diseases such as lung cancer) and the benefits of quitting with positive sentiment (e.g., healthy life) using research findings and real patient cases [190]. The news media influence the understanding of public health concerns by selecting specific aspects of a topic and presenting the concerns as salient news articles [191]. Previous research assessing sentiments and the focus of news media toward public health were conducted using traditional qualitative content analysis. Glenn et al. analyzed the sentiments

of national online news and the readers' comments for weight loss surgery [192]. They found that the sentiments of the news articles were mostly positive and supportive, while the sentiments of readers' comments were predominately negative and associated with some negative terms such as "piggy" and "fatty". Patterson et al. analyzed the content of seven UK national newspapers to identify the style of presentation of news media for women's and men's drinking [193]. Their findings indicated a difference by participants' gender. For instance, men's drinking was mostly associated with the topics of "violence" and "disorderly", while the women's drinking was frequently linked to the topics "out of control", "putting themselves in danger", "harming their physical appearance" and "burdening men". Although providing useful insights into the sentiments and focuses of news media for a specific healthcare issue, qualitative content analytic approaches are costly, time-consuming, and resource intensive. There is the subjective nature of traditional qualitative inquiry due to human perceptions and interpretations. This subjectivity does not avail itself to efficiently or systematically detecting sentiments and the focus of news media. Another key factor related to the sample sizes used in the qualitative content analysis is usually limited to a few hundred of news articles, which could limit the generalizability of the findings. To address these challenges, we used state-of-art text mining methods including sentiment analysis and topic modeling, together with statistical analysis, to efficiently analyze more than 3 million Reuters news articles to identify news coverage, sentiments, and emphases toward public health issues from 2007 to 2017. We identified 10 major public health issues (i.e., "air pollution", "alcohol drinking", "asthma," "depression", "diet", "exercise", "obesity", "pregnancy", "sexual behavior" and "smoking")

based on the top keywords from public health scientific journals. Sentiment analysis refers to the use of computerized algorithms for systemically evaluating opinions (e.g., negative or positive sentiments) and their intensities of the words and sentences in a large collection of text documents [194]. Topic modeling employs computerized algorithms to automatically discover the hidden topics in a large body of text documents related to a specific subject. The analysis of news media data with advanced text mining techniques allows the discovery of sentiments and focuses of news media for public health issues. These discoveries could shed light on the understanding of the most pressing health concerns and provide insight for public policy. Moreover, to our knowledge, no previous work used sentiment analysis and topic modeling for identifying sentiments and focuses of news media for public health issues.

7.2 Methods

Figure 31 shows a schematic view of the methods for mining Reuters news articles in this work. The methods consist of five main phases: (1) identifying the major public health issues from 30 top public health journals; (2) downloading, cleaning, and filtering news articles from Reuters news agency for the public health issues; (3) calculating the coverage of news articles over a decade linked to the public health issues and compare them with Google Trends searches; (4) analyzing sentiments of news articles related to the public health issues and their trends over time; and (5) identifying the focuses of the news articles associated with the public health issues. We briefly describe the five phases in the following subsections with detailed descriptions in Supplementary document 1.

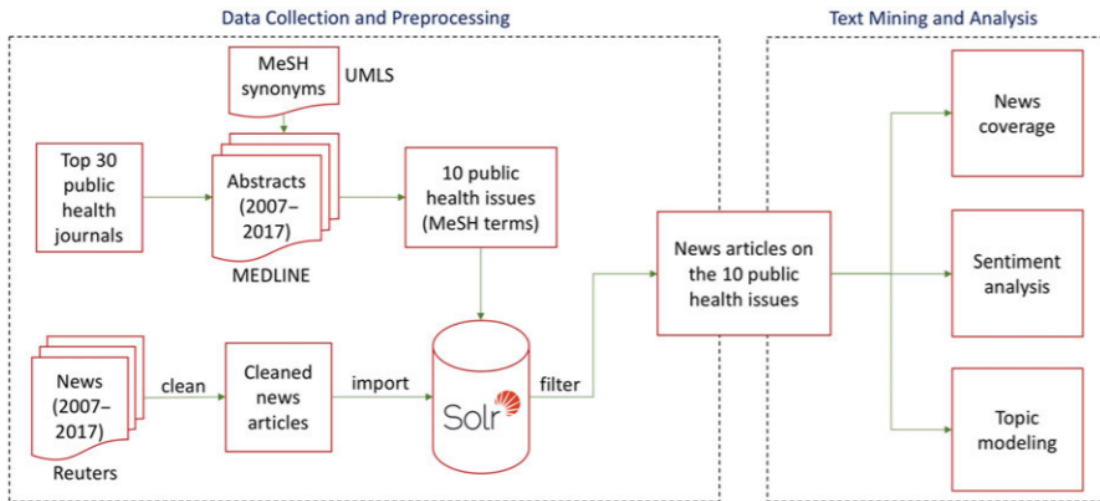


Figure 31: A schematic view of methods for mining Reuters news. MeSH, Medical Subject Heading; UMLS, Unified Medical Language System

The Python scripts for mining news articles can be accessed via Github [huang2019a](#).

Identifying Major Public Health Issues Public health research studies generally investigate major public health issues to provide new knowledge and insights to increase wellness of the general population. These studies are mainly published in public health journals that are indexed in the MEDLINE database [195]. Abstracts of the published articles including the keywords present their main research focuses. Thus, we selected 30 of the top public health journals (See Supplementary document 2) and downloaded 61,387 abstracts of articles published between January 1, 2007 and December 31, 2017 to identify the major public health issues. We mapped the article keywords to Medical Subject Headings (MeSH) terms. MeSH is a controlled terminology developed by National Library of Medicine for indexing articles in the MEDLINE database [196]. We identified the synonyms of the MeSH terms using Unified Medical Language System (UMLS)

Table 27: Frequencies of MeSH terms related to public health

Public health issue(MeSH term)	Frequency
Smoking	4936
Obesity	1403
Pregnancy	3583
Air pollution	1009
Exercise	869
Diet	836
Sexual behavior	797
Alcohol drinking	788
Depression	755
Asthma	669

Metathesaurus [197], [198]. The UMLS Metathesaurus is a collection of controlled terminologies and provides mapping structures between different medical vocabularies via concept unique identifier. Sequentially, we developed a Python script with regression expression **huang2019a** to identify the MeSH terms (and their synonyms) in the abstracts and calculated the frequencies of the MeSH terms. The frequency of a MeSH synonym is added to the frequency of the MeSH term. After removing the MeSH terms which were not related to public health or whose relative frequencies are less than 1%, we found 90 popular MeSH terms on public health (see Supplementary document 3). We selected 10 major public health issues (table 27) out of the 90 MeSH terms based on the frequencies.

7.2.0.1 Collecting, Cleaning, and Filtering News Articles

Collecting News Articles From Reuters News Agency Reuters News Agency is a leading global information media agency and the world’s largest international text and television news provider [199]. We developed a web crawler in Python **huang2019a** to

download news articles from an online archive of Reuters news agency [200]. We collected 3,763,737 articles between January 1, 2007 and December 31, 2017 to investigate major public health issues in the articles.

7.2.0.2 Cleaning News Articles

News articles contain noise and metadata that could affect the results of sentiment analysis and topic modeling toward public health issues. After reviewing a small sample of articles, we identified patterns that needed removing from the news articles. We removed repetitive special characters, such as "â", from the articles. We also removed the tags, such as "(Reuters)" and editorial information (i.e., "reporting by" or "editing by"). We replaced the sentence delimiters "*" and ";" using a period and replaced the hyperlinks, such as "http://topnews.session.com" with the word "link". In addition, readers' comments were deleted in the articles since the focus of the analysis is the content of the story reported.

Filtering News Articles on Public Health Issues To filter articles related to the 10 public health issues (27), we imported the articles into Apache Solr for information indexing and searching. Apache Solr is an open source search platform built on Apache Lucene library. Apache Lucene provides rich features to handle document such as full-text search and real-time indexing for various applications **singh-a**, [201]. The Reuters news data were filtered in Apache Solr to retrieve articles that mentioned the 10 public health issues.

7.2.0.3 News Coverage

Descriptive statistics were used to calculate the coverage of news media (i.e., numbers of news articles) for the 10 public health issues over time [202]–[205]. We compared the coverage trends of articles to Google Trends searches for the public health issues [202], [203]. Google Trends analyzed the Internet search patterns of the individuals using Google search services over time. The Internet search patterns reflect information-seeking behaviors of the individuals.

Google Trends provides the majority of the Internet search services and makes the searching data publicly available.

7.2.0.4 Sentiment Analysis

Sentiment is a view of or attitude (e.g., positive or negative opinion) toward a situation or event. Sentiment analysis denotes systematic evaluation of opinions and their intensities of the words and sentences in a large collection of text documents by using computerized algorithms [194]. Sentiment analysis is widely used in the area of health-care [194], [199], [206], [207], particularly for identifying the attitudes and opinions of patients' posts in social media toward a specific healthcare issue. For example, Hopper and Uriyo used sentiment analysis to review patients' feedback for a selected group of gynecologists in Virginia [208]. In another study, Clark et al. applied sentiment analysis to quantify sentiments of patients toward breast cancer treatment experience [209]. In our previous work, we also used sentiment analysis to identify the sentiments of news articles toward hundreds of diseases and medical conditions [199]. Computerized algorithms

were developed to automatize the process of sentiment analysis, enabling researchers to evaluate sentiments of a large volume of text documents. In this work, we used a python module, Valence Aware Dictionary and sEntiment Reasoner (VADER) [210], to quantify the sentiments of the articles toward the 10 identified public health issues. VADER was specifically tuned to identify the sentiments of a wide range of social media data [199], [211]. VADER reports a normalized and weighted sentiment score for a given sentence, according to predefined score of each word and embedded rules. The reported sentiment score is between -1.0 (the most negative) and 1.0 (the most positive), with 0.0 indicating neutral. To improve the accuracy of sentiment analysis for each public health issue, we measured the sentiments of sentences containing MeSH terms and their synonyms for the public health issue. We calculated the average of sentiment scores of all sentences related to the public health issue in articles as a sentiment score for the public health issue. We computed the average sentiment scores of all sentences linked to a public health issue in all the news articles as a sentiment score of news for the public health issue in each year. We classified an article as positive, neutral, or negative, according to the threshold values suggested by VADER [211]. More specifically, if a sentiment score of a news article is equal to or larger than 0.05 , the sentiment of the news article is positive; if a sentiment score of a news article is less than 0.05 and larger than -0.05 , the news article has a neutral sentiment; otherwise, the news article is negative.

7.2.0.5 Topic Modeling

Due to the challenges (e.g., intensive human labor) of topic analysis with traditional manual qualitative methods, topic modeling methods were developed to automatically identify hidden topics in a massive collection of text documents [212], [213]. The most frequently used topic modeling method is Latent Dirichlet Allocation (LDA) that was introduced by Blei et al. in 2003 [214]. LDA was extended and adopted in several domains for different purposes such as news themes on diseases, prognosis of human papillomavirus infection [215], and technology innovation in patents [212]. Although LDA is a powerful tool for discovering hidden topics in a large set of text documents, it is associated with some limitations. For example, LDA neglects the important word order in a text document. The text document is not treated as a sequence of words instead it is treated as a "bag" of words for topic modeling. LDA cannot automatically detect the number of topics in the text document and requires a predefined topic number as an input for topic modeling. Some posterior techniques such as perplexity and topic coherence [214], [216] could help tune and find the appropriate number of topics in the text document, and they require extra computational time and involvement of domain experts to examine the generated topics for determining a good topic number. Given the number of topics, LDA infers topic distribution in each document (e.g., $0.5 * \text{topic1} + 0.3 * \text{topic2} + 0.2 * \text{topic3}$ for the document1) and word distributions over a topic (e.g., $0.4 * \text{word1} + 0.3 * \text{word2} + 0.2 * \text{word3} + 0.1 * \text{word4}$ for the topic1). During the inference of word distribution over a topic, LDA treats a common word (i.e., a word occurs equally across all topics) and a characteristic word (i.e., a word occurs dominantly in a few topics) equally when

they have the same probability given a topic. In response to these limitations, we used an advanced topic modeling method, Topic Keyword Model (TKM) [217], [218], to identify the hidden topic structure of articles related to the 10 public health issues. TKM considers the word order in a text document for topic modeling. From a human perspective, it seems that multiple consecutive words in a text document have large probabilities to associate with the same topic. Similarly, TKM links a word to a topic if it or its adjacent words have a high association score with the topic. Thus, the topic that a word links to is heavily influenced by its nearby words. This way, the order of words was involved in topic modeling with TKM, compared to LDA. TKM measures the dissimilarity between topics and only keeps the topics that are significantly different from each other during topic modeling. TKM could potentially determine the appropriate number of distinct topics in the text document. In addition, TKM differentiates a common word and a characteristic word for a topic and adjusted the association probability (score) of a word with a topic according to its commonness and distinctiveness among topics while inferring word distributions over a topic. After word stemming and lemmatization that reduce words into their base forms [219], we used the TKM package provided by its author in GitHub [218] to learn the topics of the articles linked to each of the 10 public health issues. Because topic modeling algorithms are unsupervised learning methods and they do not require prior annotation as gold standard to train the models, the standard metrics (e.g., recall, precision, and F-measure) for supervised learning methods are not suitable for evaluating the results of topic modeling. Therefore, we asked domain experts to evaluate the results of topic modeling [220].

7.3 Results

7.3.1 Findings of News Coverage

We calculated the coverage of articles associated with each of the 10 public health issues every year between 2007 and 2017. The results are compared with the numbers of Google Trends searches for the 10 public health issues as shown in 32. We rescaled these numbers relative to the highest number on each subfigure. Figure 32 shows that the numbers of news articles for the seven public health issues "Smoking", "Exercise", "Alcohol drinking", "Diet", "Obesity", "Depression" and "Asthma" were constantly decreasing over years. The numbers of news articles for the remainder of the public health issues, "Sexual behavior", "Pregnancy" and "Air pollution" fluctuated in the study period. We found that the decreasing trends of Google searches for "Smoking" and "Obesity" are in line with the trends of relevant articles. In contrast,

the number of Google searches for "Alcohol drinking" steadily increased over time, which had a negative correlation with the number of articles for "Alcohol drinking."

7.3.2 Findings of Sentiment Analysis

Figure 33 shows the frequencies of positive, neutral, and negative sentiments of the articles toward the 10 public health issues. We found that the sentiments of the news articles on the three of the public health issues, "exercise," "alcohol drinking," and "diet" were predominately positive (i.e., 55.6%, 43.4%, and 45.6%, respectively), implying that the articles associated these issues with positive terms, such as "happiness," "energy," or terms showing overall healthy life. For "alcohol drinking," we found that there were more

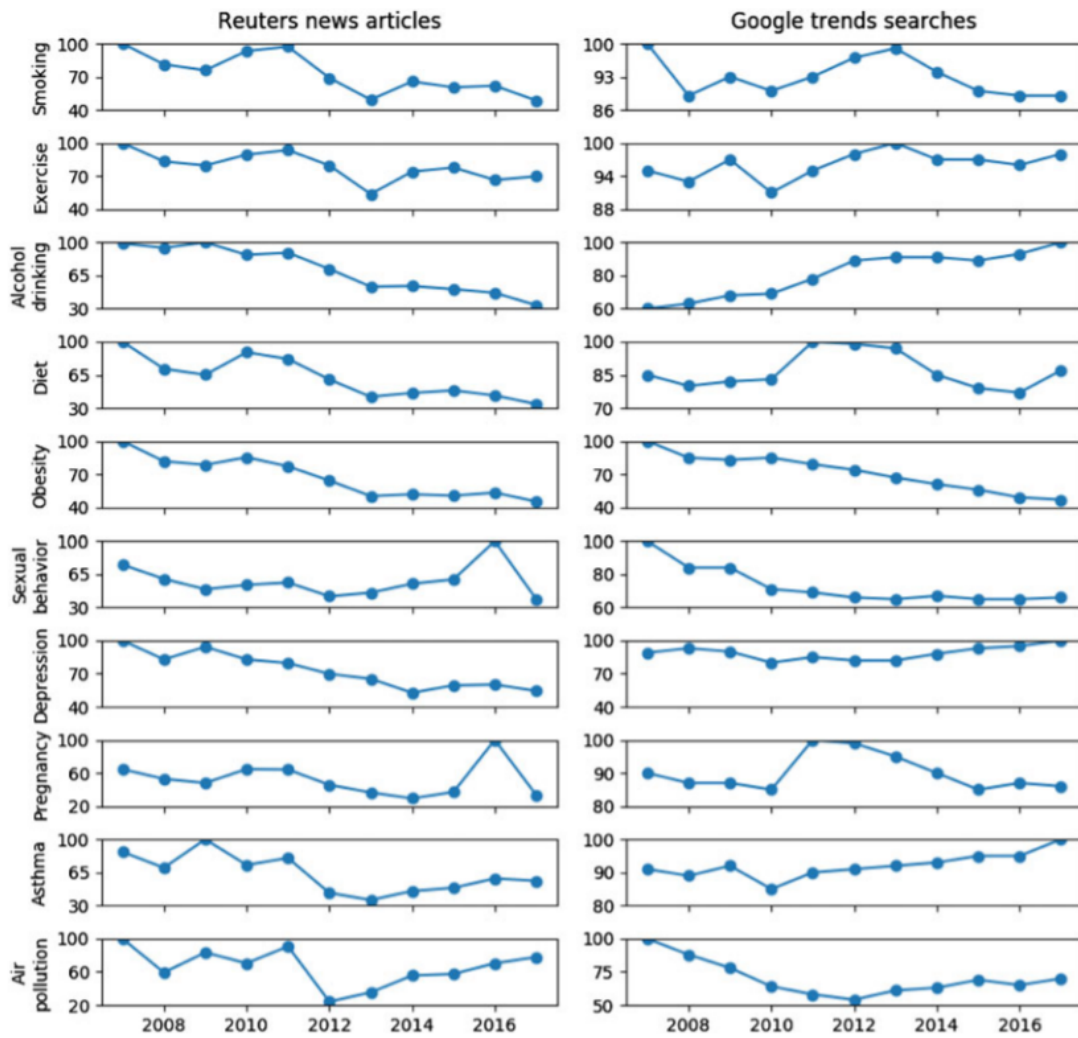


Figure 32: Normalized numbers of articles and Google Trends searches for the 10 public health issues over time. The numbers are normalized to the highest point on each subfigure. A value of 100 represents the peak popularity for the public health issue.

news articles with positive sentiment (43.4%) than news articles with negative sentiment (29.1%), which is surprising due to the public concern about alcohol misuse. For the public health issues of "smoking," "obesity," "sexual behavior," "depression," "pregnancy," "asthma," and "air pollution," Reuters published 50.0%–89.1% articles with negative sentiment and 8.9%–31.6% articles with positive sentiment. This occurred more frequently because the articles were linked the topics with negative terms, such as diseases, symptoms, low quality of life, "pressure," "hopelessness," and "worrying." In addition, we found that among these public health issues, "smoking" is the mostly mentioned in the articles and "depression" had the largest coverage percentage (89.1%) of articles with negative sentiment. Figure 34 shows the sentiment scores of the media toward the 10 public health issues over 11 years (2007–2017). During 2007–2017, "depression" had the lowest sentiment score that fluctuated between -0.6 and -0.4 , compared with other public health issues. On the other hand, "exercise" showed the highest sentiment score after 2007 and its sentiment score steadily increased between 2007 and 2017. This finding could imply that exercise was increasingly linked to terms for reducing disease risk and improving life quality in the news articles.

7.3.3 Findings of Topic Modeling

We used TKM to identify topics of articles associated with 10 public health issues. In this section, we presented the identified topics for two public health issues, "smoking" and "alcohol drinking" (Figure 35). For the topics associated with the remaining eight public health issues, please see Supplementary document 4. TKM identified 14 topics

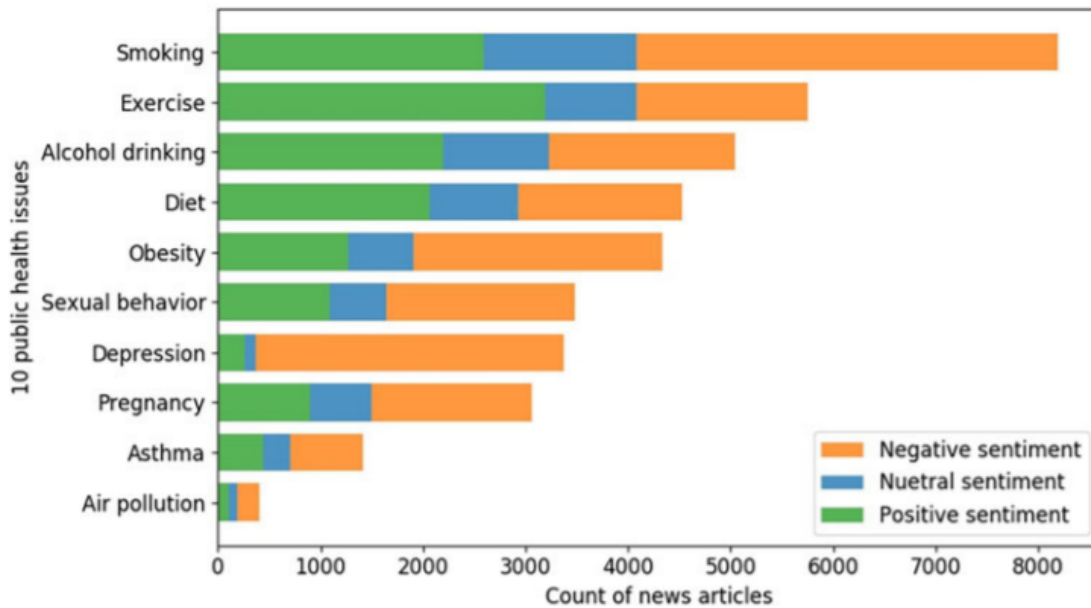


Figure 33: Counts of news articles with positive, neutral, and negative sentiments toward 10 public health issues.

of the news articles related to "smoking." We selected five most meaningful topics and showed them in Fig. 5 (see Supplementary document 4 for the rest nine topics). By interpretation of the identified topic keywords, we could find that the five meaningful topics of the news articles on "smoking" was mostly related to "tobacco and cigarette," "industry," "adolescent smoking," "cancer," and "cardiovascular disease." For articles on "alcohol drinking," TKM discovered 16 topics and we selected and illustrated five most meaningful topics in Figure 35. The remainder of the 11 topics is in Supplementary document 4. After interpretation of the identified topic keywords, we found that the five meaningful topics on "alcohol drinking" are "health research," "driving," "wine industry," "culture and diet constraint," and "opioid."

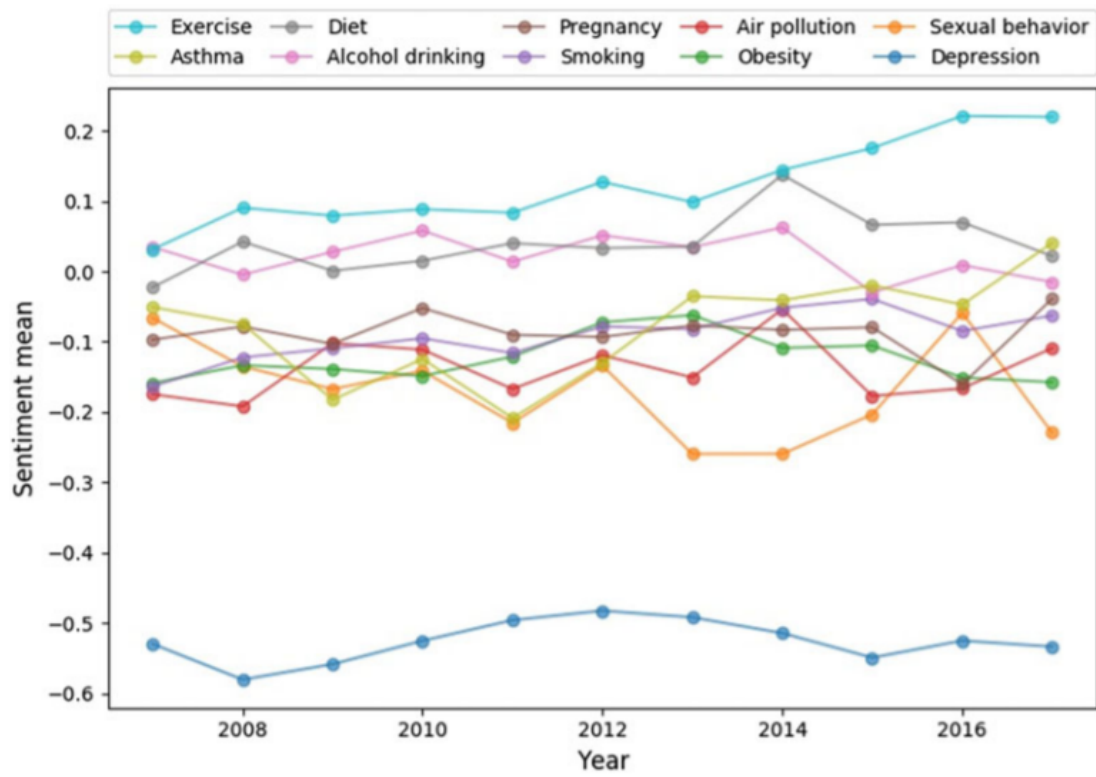


Figure 34: Sentiment scores of news media toward 10 public health issues over 11 years (2007-2017).



Figure 35: Word clouds of five meaningful topics identified in news articles related to the public health issues, "smoking" and "alcohol drinking."

7.4 Discussion

The results of sentiment analysis showed that the sentiments of over 43% news articles toward "exercise", "diet" and "alcohol drinking" were positive. The dominantly positive sentiments "exercise" and "diet" could imply that the news articles mostly focused on the importance of healthy diet and regular exercise and their relationship with disease prevention and high quality of life. For example, articles such as "Study details how high fiber diets make for healthier lives" [221] and "Aerobic exercise eases depression, even in chronically ill" [222] linked the healthy diet and exercise to a healthy lifestyle and disease treatment. There is evidence indicating that "exercise" and/or "diet" (e.g., a Mediterranean diet) serve as a preventive or disease-modifying treatment of diseases such as dementia **morris-a**, [223], Parkinson's disease [224]–[226], and cardiovascular disease [227]–[229].

These studies possibly explain the constant increase of sentiment score of news media toward "exercise" over time. It seems to be surprising that the majority of the articles on "alcohol drinking" has positive sentiment, due to the public concern about excessive alcohol use. However, news articles such as "Drinking alcohol may keep leg arteries healthy" [230] and "Moderate drinkers have a better health, study finds" [231] showed how articles had a positive sentiment by highlighting the positive aspects of "alcohol drinking".

The attitude of articles toward "alcohol drinking" supported the findings by Mostofsky et al. [232]. According to their study, moderate alcohol drinking was associated with

lower risk of cognitive decline and heart diseases. But moderate or higher alcohol intakes increase the risk of diseases such as breast cancer and bone fracture, particularly in women. On the other hand, the sentiments of more than 50% articles were negative for "smoking", "obesity", "sexual behavior", "depression", "pregnancy", "asthma", and "air pollution". The predominantly negative sentiments for these public health issues could be justified in two different ways: (1) the potential link of the public health issues to diseases, low work productivity, or poor quality of life. For example, articles such as "Air pollution a leading cause of cancer" [233], "Poor mental health harming productivity, says OECD" [234], and "Sexual harassment, abuse tied to real health effects" [235] linked "air pollution," "mental disorder" (e.g., depression), and "inappropriate sexual behavior" to cancer, low productivity, and side effects on health, respectively. (2) The healthcare services and public health policies and programs are not effective enough to manage the issues. For example, articles such as "Kids with asthma often leave doctor's office with unanswered questions" [236], "Obesity medical bill could reach 1.2 trillion a year by 2025" [237], and "Pot-smoking on the rise among U.S. pregnant women" [237] indicate the ineffectiveness of public health and healthcare interventions for managing "asthma," "obesity" and "pregnancy" respectively. This is particularly the case for "depression" with the lowest sentiment score from the media, because it is currently one of the major public health concerns in our society and it is associated with several limitations in daily functioning and social participation [238].

We identified the main topics of the articles on the 10 public health issues using topic modeling. The highlighted five meaningful topics for two public health issues, "smoking" and "alcohol drinking". The five major topics emphasized in the articles related to "smoking" were "tobacco and cigarette" "industry" "cancer" "cardiovascular disease" and "adolescent smoking". The topics "cancer" and "cardiovascular disease" indicate that these fatal diseases were strongly associated with smoking in the media. It is well documented that smoking is linked to more than 12 cancers, particularly, lung cancer [239]. About 90% lung cancers are caused by tobacco smoking or secondhand smoke exposure **health2014b**. Smoking is linked to cardiovascular health and cardiovascular disease with smoking attributing to about 140,000 premature deaths annually. The topic "adolescent smoking" suggests that tobacco use in adolescence is another focus of the articles. In 2013, about 18% middle school students and 46% high school students were tobacco users [240]. Nearly 90% cigarette smokers first tried cigarette smoking by age 18, and the prevention of tobacco use in adolescence is critical to reducing tobacco epidemic in the US [241]. The five major topics uncovered in the articles on "alcohol drinking" are "health research", "driving", "wine industry", "culture and diet constraint" and "opioid". The topic "health research" may show that health-related research on "alcohol drinking" was mainly discussed in the articles, for example, the news articles "Cutting back on alcohol can prevent cancers: experts" [242] and "Moderate drinking helps heart, but don't binge" [243]. The topic "opioid" suggests the mixed use of opioid and alcohol and related risks. The opioids are effective painkiller but have the potential to be additive [244]. The combined use of opioid and alcohol increases the risk of overdose and injury [245]. The

topic "driving" implies that the alcohol impaired driving was another focus of many of the articles, possibly because driving under the influence of alcohol remains a public health problem [246].

The use of big data to understand public health issues is a novel way to provide clinical and translational science awards programs insight on community priorities that lend themselves to community-engaged research approaches [247]. Our other research indicates that meaningful community engagement offers the opportunity to promote bi-directional dialogue about health research with diverse communities [247]–[249]. The process used in this study offers a new way to identify topics for future dialogs with community-engaged stakeholders to set research priorities.

The discussion of this study cannot be considered fully without analyzing its limitations. The first limitation is related to the data source. We used articles published by Reuters between 2007 and 2017 for studying major public health issues because Reuters is a leading news media organization and the largest international text news provider. Although we are confident that the findings are not atypical for other media sources that cover public health issues, the results of coverage, sentiment analysis and topic modeling might not be generalized to other news media agencies (e.g., Associated Press) or other forms of media (e.g., television, radio or social media). In addition, the findings of news media may not reflect the concerns of the US population which is another indication that the findings lend themselves for starting a dialog about these topics as potential areas of focus for health-related research with the community.

The second limitation of the study is the synonyms generated using UMLS for

each public health issue. Since UMLS is a compendium of standard medical terminologies, it might not include all synonyms that Reuters journalists use when writing about public health issues. The third limitation relates to the sentiment analysis method (VADER) used for identifying sentiments of articles toward the public health issues. Although VADER has been evaluated using articles published by New York Times, it has not been tested or tuned on news articles from Reuters. The fourth limitation is about the topic modeling method. We used TKM to identify topics in news articles for each public health issue. Although TKM addresses the limitations of LDA, topic modeling with TKM was performed in a completely unsupervised fashion. To evaluate the results of topic modeling, we relied on domain experts' judgments that might create bias in interpretation of these topics.

7.5 Conclusion

In this study, we identified 10 important public health issues after analyzing more than 60,000 abstracts of 30 top public health journals. We analyzed over 3 million Reuters articles during 2007â2017 identified their sentiments with linkages to the 10 public health issues, using state-of-art text mining methods including sentiment analysis and topic modeling. Our results show that the coverage of news articles associated with each of the seven public health issues, "Smoking," "Exercise," "Alcohol drinking," "Diet," "Obesity," "Depression," and "Asthma" had a declining trend over years. The coverage of news articles for the rest three public health issues, "Sexual behavior," "Pregnancy," and "Air pollution" fluctuated over time. For sentiment analysis, the sentiments of the news articles for the

three public health issues, "exercise," "alcohol drinking," and "diet," were predominately positive. It suggests that the articles associated these issues with positive terms such as energy. For the remainder of the seven public health issues including "smoking," "obesity," "sexual behavior," "depression," "pregnancy," "asthma," and "air pollution," most articles had negative sentiments. It may indicate that the articles mostly linked these issues to negative terms such as diseases or symptoms. Our study showed that text mining methods may address the limitations associated with traditional qualitative approaches. Our analysis could provide valuable insights about the sentiments and topic structures of articles discussing public health issues. Our findings could offer valuable information for the healthcare professionals and policy makers.

CHAPTER 8

CONCLUSION

We proposed a novel framework for understanding the textual data based on the competitive learning. We applied three different approach 1) Second Chance Autoencoder for Textual data (SCAT) 2) similarity Based SCAT (SSCAT) and 2) Coherence Based SCAT (CSCAT) which works based on similarity and coherence score respectively.

8.1 Second Chance Autoencoders For Textual Data

- SCAT and SSCAT: As shown in Chapter 2, SCAT and SSCAT are two innovative autoencoders for textual data that we proposed. The models are based on the concept of k -competitive learning, in which a significant subset of k winning neurons engage in the learning process while the remaining neurons remain inactive. The winning neurons become highly specialized in learning distinctive features as a result of competition. Unlike prior techniques, which encouraged competition between the most powerful positive and negative neurons, our strategy begins by removing extremely comparable neurons. It then gives a completion for the autoencoder's bottleneck layer's top and lowest, positive and negative neurons.

Our investigations shown that our technique delivers very similar or superior performance outcomes in a variety of textual data applications, such as classification, topic modeling, and document visualization. Additionally, our models encompass a

greater number of semantically critical subjects than the baseline models reviewed in this work. As a result, the suggested approach is well-suited for textual data dimension reduction.

- **Coherence Based SCAT (CSCAT):**

As shown in Chapter 3, CSCAT is a novel textual data autoencoder that we suggested. Unlike previous neural network-based models, CSCAT may develop meaningful representations for textual content by intentionally driving the model to be sparse. We demonstrate a considerable increase in both F1 score/precision/recall accuracy and recall accuracy when compared to previous topic models. Additionally, the topic coherence is greater than that of the SSCAT model, which highlights the fact the model's architecture, which promotes neurons to maintain highly coherent features. Interestingly, despite the fact that we utilize a shallow model, with only one hidden layer, it beats a number of other approaches on a wide variety of text analytics tasks. We compare CSCAT against a variety of approaches, including graphical models (e.g., LDA), W-LDA, KATE, and numerous other autoencoders. Across tasks such as document classification, multi-label classification, and coherence, we show that CSCAT consistently outperforms competitive approaches or achieves results that are near to the best. It is extremely promising to observe that CSCAT may also acquire semantically meaningful representations of words, documents, and topics, as demonstrated by quantitative and qualitative investigations.

8.2 Application Of The Topic Modeling In Different Domain

- Early detection of Alzheimers' diseases: As shown in Chapter 6 ??, The health-care sector has historically been an early user of new technologies and has profited significantly from them. Machine learning now plays a critical role in a variety of health-related fields, including the creation of novel medical procedures, the management of patient data and records, and the treatment of chronic diseases. One of the aspect that machine learning has been used is in detecting the disease. However, early detection of the disease has a significant effect on the health case since it can prohibit financial problems and also save many people people. There are significant variations in the temporal trends of b-ADL and i-ADL between CI and CU patients roughly 1 to 1.5 year(s) before physician-diagnosed CIâi.e., a steeper slope of total ADL deterioration and poorer i-ADL than b-ADL in CI patients during this period. Individual ADL (bathing and responsibility for one's own medicine) trajectories were found to be strongly linked with the development of CI. The themes and phrases extracted from clinical free text using topic modeling techniques have the potential to demonstrate how CI patients' problems change over time and to disclose previously unrecognized conditions as they approach CI diagnosis. These findings may contribute to the early diagnosis of CI and consequently to the expedited treatment of underlying disorders and concomitant problems. Around 1 to 1.5 year(s) before to the actual physician diagnosis of CI, the trajectories of ADL deterioration increased steeper in CI patients than in CU patients. The topic modeling revealed that the majority of the topic words were associated and accurately

described the underlying semantics of CI when approaching CI diagnosis. In the future, we intend to use neuroimaging and assessment data to refine the categorization of cognitive function and to construct a prediction model based on our observations to identify patients at high risk of developing different stages of CI and to uncover related longitudinal risk factors.

- Analyzing public health: In Chapter 7, we explored the potential use of topic modeling approaches in public health-related discoveries and public health surveillance. We employed descriptive statistics and state-of-the-art text mining to conduct sentiment analysis and topic modeling on over 3 million Reuters news stories published between 2007 and 2017 in order to determine their coverage, sentiments, and focus on health and safety concerns. The news coverage of seven public health concerns has dropped over time: "Smoking," "Exercise," "Alcohol use," "Diet," "Obesity," "Depression," and "Asthma." Between 2007 and 2017, press coverage of "Sexual conduct," "Pregnancy," and "Air pollution" varied. The sentiments expressed in news stories about three public health concerns, "exercise," "alcohol use," and "diet," were mostly favorable and included words like "energy." The next seven public health concerns elicited predominantly negative responses, which were associated with negative phrases, such as illnesses. The topic modeling findings mirrored the media's emphasis on public health problems.

Text mining techniques have the potential to overcome the constraints of traditional qualitative methodologies. Utilizing big data to better understand public health needs is an innovative strategy that might assist clinical and translational science

award programs in concentrating their efforts on community-engaged research activities that address community goals.

- Personality detection in social media: In Chapter 4, we studied the use of machine learning based algorithms for identifying the personality of users along with challenges in this domain. The exponential expansion of social media users has resulted in a huge increase in the volume of online content. Often, the contents that these users post on social media can provide valuable insight into their personalities without requiring them to take formal personality tests. The approach, called personality prediction, entails segmenting digital input into elements and mapping them to a personality model. Due to its simplicity and demonstrated competence, a well-known personality model known as the big five personality characteristics has frequently been embraced as the de facto benchmark for personality evaluation in the literature. This study analyzes several feature extraction techniques and algorithmic approaches for developing a personality prediction system using data from diverse social media sources. In this experiment, the suggested deep learning architecture technique shown that NLP statistical characteristics surpass the majority of personality model builds in terms of accuracy. Additionally, the personality prediction system benefited from NLP statistical characteristics such as the TF-IDF and LIWC lexicon database. Additionally, from a construct validity standpoint, we can demonstrate that a machine-learned model is theoretically more fit for labeling social media writing than the Personality Recognizer, which was built using monologues.

- Opioid crisis in social media: As discussed in 5 opioid abuse has long been seen as a severe risk to public health in the United States (Centers for Disease Control and Prevention, 2017). We can forecast how the abuse of opioid has been presented in social media and how it can impact people’s opinions using the latest machine learning-based algorithm and data published by users on social media. Given the patterns of topics, sentiment, and time series, public discourses on social media about the opioid epidemic captured immediate replies directly from regular Twitter users. To average consumers, the opioid epidemic has resulted in a staggering number of fatalities and catastrophes; it is a severe public health problem that must be addressed and resolved. In general, social media agendas can respond to crises faster than traditional news reporting and offer a broader range of opinions. Thus, studying public discourses on social media can assist in identifying early warning indications of public concern about health issues.

8.3 FUTURE WORKS

Our future work aims to extend "Second Chance" concept to other neural network architectures, including variational autoencoders, GAN, CNN, etc. We want to explore how the idea of giving a second chance to the neurons with small activation values can benefit the model and reveal more significant patterns or help in classification/clustering approaches. Another exciting area to explore is to apply this concept to image data sources.

We also plan to consider the density of features corresponding to neurons to filter

out uninteresting features if they are not deterministic of the corresponding latent variable. That is, we select the eligible neurons based on the intensity of neurons; the intensity can be calculated by taking the reverse of entropy.

Also, another interesting thing is to look at the statistics of the data and then choose which method to apply. For example, if the distribution of the data in our corpus is skewed towards low frequent words, it shows that we have many tokens, that their frequency is low, and they may be dominated by top 20 percent high frequent tokens. In this scenario, we may decide to give more energy to the lowest weight neurons against neurons with the highest activation values. Thus, gaining knowledge about the data statistics will help design more dynamic models such that various parts of the architecture will be activated accordingly based on the input data distribution.

Bibliography

- [1] A. Agrawal, W. Fu, and T. Menzies, “What is wrong with topic modeling? and how to fix it using search-based software engineering,” *Inform Software Tech*, vol. 98, pp. 74–88, Jun. 2018.
- [2] X. Yan, J. Guo, Y. Lan, and X. Cheng, “A biterm topic model for short texts,” in *Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 2013*, pp. 1445–1455.
- [3] D. Blei and J. Lafferty, “Dynamic topic models,” in *Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 2006*, pp. 113–120.
- [4] J. Lafferty and D. Blei, “Correlated topic models,” in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Scholkopf, and J. Platt, Eds., MIT Press, 2006, pp. 147–154.
- [5] W. Li and A. McCallum, “Pachinko allocation: Dag-structured mixture models of topic correlations,” in *Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 2006*, pp. 577–584.
- [6] X. Wang, Y. Zhao, and F. Pourpanah, *Recent advances in deep learning*, 2020.
- [7] M. A. Wani, F. A. Bhat, S. Afzal, and A. I. Khan, *Advances in deep learning*. Springer, 2020, vol. 57.

- [8] N. O’Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh, “Deep learning vs. traditional computer vision,” in *Science and Information Conference*, Springer, 2019, pp. 128–144.
- [9] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller, “Deep learning for environmentally robust speech recognition: An overview of recent developments,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 9, no. 5, pp. 1–28, 2018.
- [10] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, “Deep learning for extreme multi-label text classification,” in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 115–124.
- [11] X. Wei and W. B. Croft, “Lda-based document models for ad-hoc retrieval,” in *Proceedings of the 29th annual international ACM SIGIR Conference on Research and development in information retrieval*, 2006, pp. 178–185.
- [12] S. Goudarzvand, J. S. Sauver, M. M. Mielke, P. Y. Takahashi, Y. Lee, and S. Sohn, “Early temporal characteristics of elderly patient cognitive impairment in electronic health records,” *BMC Medical Informatics and Decision Making*, vol. 19, no. 4, p. 149, 2019.
- [13] S. Goudarzvand, J. S. Sauver, M. M. Mielke, P. Y. Takahashi, and S. Sohn, “Analyzing early signals of older adult cognitive impairment in electronic health records,” in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2018, pp. 1636–1640.
- [14] S. Fouladvand, M. M. Mielke, M. Vassilaki, J. S. Sauver, R. C. Petersen, and S. Sohn, “Deep learning prediction of mild cognitive impairment using electronic

- health records,” in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2019, pp. 799–806.
- [15] M. Hosseini, A. S. Maida, M. Hosseini, and G. Raju, “Inception lstm for next-frame video prediction (student abstract),” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 13 809–13 810.
- [16] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, “Greedy layer-wise training of deep networks,” in *Advances in Neural Information Processing Systems*, 2007, pp. 153–160.
- [17] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [18] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, p. 436, 2015.
- [19] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.
- [20] Z. Zhu, X. Wang, S. Bai, C. Yao, and X. Bai, “Deep learning representation using autoencoder for 3d shape retrieval,” *Neurocomputing*, vol. 204, pp. 41–50, 2016.
- [21] Y. Chen and M. J. Zaki, “Kate: K-competitive autoencoder for text,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2017, pp. 85–94.
- [22] S. Zhai and Z. M. Zhang, “Semisupervised autoencoder for sentiment analysis,” in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 993–1022.

- [23] H. Larochelle and S. Lauly, “A neural autoregressive topic model,” in *Advances in Neural Information Processing Systems*, 2012, pp. 2708–2716.
- [24] L. Maaloe, M. Arngren, and O. Winther, “Deep belief nets for topic modeling,” *arXiv preprint arXiv:1501.04325*, 2015.
- [25] Y. Miao, L. Yu, and P. Blunsom, “Neural variational inference for text processing,” in *International Conference on Machine Learning*, 2016, pp. 1727–1736.
- [26] C. Zhang, J. Butepage, H. Kjellstrom, and S. Mandt, “Advances in variational inference,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
- [27] A. Makhzani and B. Frey, “K-sparse autoencoders,” *arXiv preprint arXiv:1312.5663*, 2013.
- [28] D. M. Blei, “Probabilistic topic models,” *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.
- [29] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, “An overview of topic modeling and its current applications in bioinformatics,” *Springerplus*, vol. 5, no. 1, Sep. 2016.
- [30] L. Hong and B. D. Davison, “Empirical study of topic modeling in twitter,” in *Proceedings of the First Workshop on Social Media Analytics*, acm, 2010, pp. 80–88.
- [31] P. G. Y. Girdhar and G. Dudek, “Autonomous adaptive underwater exploration using online topic modeling,” in *Experimental Robotics: The 13th International Symposium on Experimental Robotics*, Heidelberg: Springer International Publishing, 2013, pp. 789–802.

- [32] N. Andrews, K. Yogeeswaran, M. Wang, K. Nash, D. Hawi, and C. Sibley, “Is social media use changing who we are? examining the bidirectional relationship between personality and social media use,” *Cyberpsychology, Behavior, and Social Networking*, vol. 1;23(11):752-60, Nov. 2020.
- [33] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003.
- [34] A. Srivastava and C. Sutton, “Autoencoding variational inference for topic models,” *arXiv preprint arXiv:1703.01488*, 2017.
- [35] L. Liu, H. Huang, Y. Gao, Y. Zhang, and X. Wei, “Neural variational correlated topic modeling,” in *The World Wide Web Conference*, 2019, pp. 1142–1152.
- [36] F. Bianchi, S. Terragni, D. Hovy, D. Nozza, and E. Fersini, “Cross-lingual contextualized topic models with zero-shot learning,” *arXiv preprint arXiv:2004.07737*, 2020.
- [37] D. M. Blei, T. L. Griffiths, and M. I. Jordan, “The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies,” *Journal of the ACM (JACM)*, vol. 57, no. 2, p. 7, 2010.
- [38] J. Zhu and E. P. Xing, “Sparse topical coding,” *arXiv preprint arXiv:1202.3778*, 2012.
- [39] J. Eisenstein, A. Ahmed, and E. P. Xing, “Sparse additive generative models of text,” in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, Citeseer, 2011, pp. 1041–1048.
- [40] K. Canini, L. Shi, and T. Griffiths, “Online inference of topics with latent dirichlet allocation,” in *Artificial Intelligence and Statistics*, 2009, pp. 65–72.

- [41] M. Bahrani and H. Sameti, “A new bigram-plsa language model for speech recognition,” *EURASIP Journal on Advances in Signal Processing*, vol. 2010, no. 1, p. 308 437, 2010.
- [42] J. Schneider and M. Vlachos, “Topic modeling based on keywords and context,” in *Proceedings of the 2018 SIAM International Conference on Data Mining*, SIAM, 2018, pp. 369–377.
- [43] T. Hofmann, “Unsupervised learning by probabilistic latent semantic analysis,” *Machine Learning*, vol. 42, no. 1-2, pp. 177–196, 2001.
- [44] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” in *Interspeech*, vol. 2013, 2013, pp. 436–440.
- [45] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [46] D. Zhou, X. Hu, and R. Wang, “Neural topic modeling by incorporating document relationship graph,” *arXiv preprint arXiv:2009.13972*, 2020.
- [47] L. Yang, F. Wu, J. Gu, C. Wang, X. Cao, D. Jin, and Y. Guo, “Graph attention topic modeling network,” in *Proceedings of The Web Conference 2020*, 2020, pp. 144–154.
- [48] Y. Bengio, *Learning Deep Architectures for AI*. Now Publishers Inc, 2009.
- [49] X. Wang and Y. Yang, “Neural topic model with attention for supervised learning,” in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2020, pp. 1147–1156.

- [50] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *arXiv preprint arXiv:1908.10084*, 2019.
- [51] A. B. Dieng, F. J. Ruiz, and D. M. Blei, “Topic modeling in embedding spaces,” *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020.
- [52] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, “Wasserstein auto-encoders,” *arXiv preprint arXiv:1711.01558*, 2017.
- [53] F. Nan, R. Ding, R. Nallapati, and B. Xiang, “Topic modeling with wasserstein autoencoders,” *arXiv preprint arXiv:1907.12374*, 2019.
- [54] P. K. Rubenstein, B. Schoelkopf, and I. Tolstikhin, “On the latent space of wasserstein auto-encoders,” *arXiv preprint arXiv:1802.03761*, 2018.
- [55] R. Wang, D. Zhou, and Y. He, “Atm: Adversarial-neural topic model,” *Information Processing & Management*, vol. 56, no. 6, p. 102 098, 2019.
- [56] A. Mingorance-Le Meur, “Jnk gives axons a second chance,” *Journal of Neuroscience*, vol. 26, no. 47, pp. 12 104–12 105, 2006.
- [57] H. Jiang and Y. Rao, “Axon formation: Fate versus growth,” *Nature Neuroscience*, vol. 8, no. 5, pp. 544–546, 2005.
- [58] S. Goudarzvand, G. Gharibi, and Y. Lee, “Scat: Second chance autoencoder for textual data,” *arXiv preprint arXiv:2005.06632*, 2020.
- [59] F. Chollet *et al.* (2015). Keras, [Online]. Available: <https://github.com/fchollet/keras>.

- [60] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, and R. Jozefowicz, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: <https://www.tensorflow.org/>.
- [61] G. Gharibi, V. Walunj, R. Alanazi, S. Rella, and Y. Lee, “Automated management of deep learning experiments,” in *Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning*, 2019, pp. 1–4.
- [62] G. Gharibi, V. Walunj, S. Rella, and Y. Lee, “Modelkb: Towards automated management of the modeling lifecycle in deep learning,” in *2019 IEEE/ACM 7th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE)*, IEEE, 2019, pp. 28–34.
- [63] M. Kuhn, K. Johnson, *et al.*, *Applied predictive modeling*. Springer, 2013, vol. 26.
- [64] K. Lang, “Newsweeder: Learning to filter netnews,” in *Machine Learning Proceedings 1995*, Elsevier, 1995, pp. 331–339.
- [65] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, “Rcv1: A new benchmark collection for text categorization research,” *Journal of Machine Learning Research*, vol. 5, no. Apr, pp. 361–397, 2004.
- [66] A. Zubiaga, “Enhancing navigation on wikipedia with social tags,” *arXiv preprint arXiv:1202.5469*, 2012.
- [67] Y. Xu and R. Goodacre, “On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating

- the generalization performance of supervised learning,” *Journal of Analysis and Testing*, vol. 2, no. 3, pp. 249–262, 2018.
- [68] V. G. Biju and C. Prashanth, “Friedman and wilcoxon evaluations comparing svm, bagging, boosting, k-nn and decision tree classifiers,” *Journal of Applied Computer Science Methods*, vol. 9, 2017.
- [69] A. Benavoli, G. Corani, F. Mangili, M. Zaffalon, and F. Ruggeri, “A bayesian wilcoxon signed-rank test based on the dirichlet process,” in *International Conference on Machine Learning*, PMLR, 2014, pp. 1026–1034.
- [70] J. H. Lau, D. Newman, and T. Baldwin, “Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality,” in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, pp. 530–539.
- [71] L. v. d. Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
- [72] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, “Adversarial autoencoders,” *arXiv preprint arXiv:1511.05644*, 2015.
- [73] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski, “Workshop on computational personality recognition: Shared task,” in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.
- [74] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, “Using linguistic cues for the automatic recognition of personality in conversation and text,” *Journal of Artificial Intelligence Research*, vol. 30, pp. 457–500, 2007.

- [75] R. V. Kozinets, “E-tribalized marketing?: The strategic implications of virtual communities of consumption,” *European Management Journal*, vol. 17, no. 3, pp. 252–264, 1999.
- [76] A. M. Kaplan and M. Haenlein, “Users of the world, unite! the challenges and opportunities of social media,” *Business Horizons*, vol. 53, no. 1, pp. 59–68, 2010.
- [77] G. Seidman, “Self-presentation and belonging on facebook: How personality influences social media use and motivations,” *Personality and Individual Differences*, vol. 54, no. 3, pp. 402–407, 2013.
- [78] M. A. DeVito, J. Birnholtz, J. T. Hancock, M. French, and S. Liu, “How people form folk theories of social media feeds and what it means for how we study self-presentation,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–12.
- [79] D. J. Stillwell and M. Kosinski, “Mypersonality project: Example of successful utilization of online social networks for large-scale social research,” *American Psychologist*, vol. 59, no. 2, pp. 93–104, 2004.
- [80] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Personality predictions based on user behavior on the facebook social media platform,” *IEEE Access*, vol. 6, pp. 61 959–61 969, 2018.
- [81] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [82] X. Ma, P. Xu, Z. Wang, R. Nallapati, and B. Xiang, “Universal text representation from bert: An empirical study,” *arXiv preprint arXiv:1910.07973*, 2019.

- [83] Y. Mehta, S. Fatehi, A. Kazameini, C. Stachl, E. Cambria, and S. Eetemadi, “Bottom-up and top-down: Predicting personality with psycholinguistic and language model features,” in *2020 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2020, pp. 1184–1189.
- [84] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio, “Twitpersonality: Computing personality traits from tweets using word embeddings and supervised learning,” *Information*, vol. 9, no. 5, p. 127, 2018.
- [85] M. P. Kalghatgi, M. Ramannavar, and N. S. Sidnal, “A neural network approach to personality prediction based on the big-five model,” *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, vol. 2, no. 8, pp. 56–63, 2015.
- [86] T. Tandra, D. Suhartono, R. Wongso, Y. L. Prasetio, *et al.*, “Personality prediction system from facebook users,” *Procedia Computer Science*, vol. 116, pp. 604–611, 2017.
- [87] S. Han, H. Huang, and Y. Tang, “Knowledge of words: An interpretable approach for personality recognition from social media,” *Knowledge-Based Systems*, p. 105 550, 2020.
- [88] J. Yu and K. Markov, “Deep learning based personality recognition from facebook status updates,” in *2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST)*, IEEE, 2017, pp. 383–387.
- [89] J. W. Pennebaker, M. E. Francis, and R. J. Booth, “Linguistic inquiry and word count: Liwc 2001,” *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

- [90] J. W. Pennebaker and L. A. King, “Linguistic styles: Language use as an individual difference.,” *Journal of Personality and Social Psychology*, vol. 77, no. 6, p. 1296, 1999.
- [91] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, “Predicting personality from twitter,” in *2011 IEEE third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE third International Conference on Social Computing*, IEEE, 2011, pp. 149–156.
- [92] Z. Lu, Y. Zhu, S. J. Pan, E. W. Xiang, Y. Wang, and Q. Yang, “Source free transfer learning for text classification,” in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.
- [93] S. J. Pan, Q. Yang, *et al.*, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
- [94] K. Clark, M.-T. Luong, U. Khandelwal, C. D. Manning, and Q. V. Le, “Bam! born-again multi-task networks for natural language understanding,” *arXiv preprint arXiv:1907.04829*, 2019.
- [95] P. Izsak, S. Guskin, and M. Wasserblat, “Training compact models for low resource entity tagging using pre-trained language models,” *arXiv preprint arXiv:1910.06294*, 2019.
- [96] G. T. Smith, “On construct validity: Issues of method and measurement.,” *Psychological Assessment*, vol. 17, no. 4, p. 396, 2005.
- [97] C. J. Hutto and E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Eighth International AAAI Conference on weblogs and social media*, 2014.

- [98] Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli, and D. Stillwell, “Personality and patterns of facebook usage,” in *Proceedings of the 4th Annual ACM web Science Conference*, 2012, pp. 24–32.
- [99] M. R. Mehl, J. W. Pennebaker, D. M. Crow, J. Dabbs, and J. H. Price, “The electronically activated recorder (ear): A device for sampling naturalistic daily activities and conversations,” *Behavior Research Methods, Instruments, & Computers*, vol. 33, no. 4, pp. 517–523, 2001.
- [100] Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli, and D. Stillwell, “Personality and patterns of facebook usage,” in *Proceedings of the 4th Annual ACM web science conference*, 2012, pp. 24–32.
- [101] G. Farnadi, S. Zoghbi, M.-F. Moens, and M. De Cock, “Recognising personality traits using facebook status updates,” in *Proceedings of the Workshop on Computational Personality Recognition (WCPR13) at the 7th International AAI Conference on Weblogs and Social Media (ICWSM13)*, AAAI, 2013, pp. 14–18.
- [102] H. Zheng and C. Wu, “Predicting personality using facebook status based on semi-supervised learning,” in *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, 2019, pp. 59–64.
- [103] M. Chen, K. Q. Weinberger, and Y. Chen, “Automatic feature decomposition for single view co-training,” in *International Conference in Machine Learning*, 2011.
- [104] D. Yao, J. Bi, J. Huang, and J. Zhu, “A word distributed representation based framework for large-scale short text classification,” in *2015 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2015, pp. 1–7.

- [105] Y. Li, A. Kazameini, Y. Mehta, and E. Cambria, “Multitask learning for emotion and personality detection,” *arXiv preprint arXiv:2101.02346*, 2021.
- [106] K. Jaidka, S. C. Guntuku, and L. H. Ungar, “Facebook versus twitter: Differences in self-disclosure and trait prediction,” in *Twelfth International AAAI Conference on Web and Social Media*, 2018.
- [107] R. Ohlsson, “Public discourse on mental health and psychiatry: Representations in swedish newspapers,” *Health*, vol. 22, no. 3, pp. 298–314, 2018.
- [108] M. N. Sellers, “Ideals of public discourse,” in *Republican Legal Theory*, Springer, 2003, pp. 62–70.
- [109] A. R. D’Agostino, A. R. Optican, S. J. Sowles, M. J. Krauss, K. E. Lee, and P. A. Cavazos-Rehg, “Social networking online to recover from opioid use disorder: A study of community interactions,” *Drug and Alcohol Dependence*, vol. 181, pp. 5–10, 2017.
- [110] E. E. McGinty, E. M. Stone, A. Kennedy-Hendricks, K. Sanders, A. Beacham, and C. L. Barry, “Us news media coverage of solutions to the opioid crisis, 2013-2017,” *Preventive Medicine*, vol. 126, p. 105 771, 2019.
- [111] X. Ji, S. A. Chun, Z. Wei, and J. Geller, “Twitter sentiment classification for measuring public health concerns,” *Social Network Analysis and Mining*, vol. 5, no. 1, p. 13, 2015.
- [112] D. M. Kotliar, “Depression narratives in blogs: A collaborative quest for coherence,” *Qualitative Health Research*, vol. 26, no. 9, pp. 1203–1215, 2016.

- [113] W. Russell Neuman, L. Guggenheim, S. Mo Jang, and S. Y. Bae, “The dynamics of public attention: Agenda-setting theory meets big data,” *Journal of Communication*, vol. 64, no. 2, pp. 193–214, 2014.
- [114] T. K. Sell, C. Watson, D. Meyer, M. Kronk, S. Ravi, L. E. Pechta, K. M. Lubell, and D. A. Rose, “Frequency of risk-related news media messages in 2016 coverage of zika virus,” *Risk Analysis*, vol. 38, no. 12, pp. 2514–2524, 2018.
- [115] S. E. Gollust, E. F. Fowler, and J. Niederdeppe, “Television news coverage of public health issues and implications for public health policy and practice,” *Annual Review of Public Health*, vol. 40, pp. 167–185, 2019.
- [116] Y. Wang and E. Willis, “Examining theory-based behavior-change constructs, social interaction, and sociability features of the weight watchers’ online community,” *Health Education & Behavior*, vol. 43, no. 6, pp. 656–664, 2016.
- [117] W. Ye and W. Erin, “Supporting self-efficacy through interactive discussion in online communities of weight loss,” *Journal of Health Psychology*, vol. 23, no. 10, pp. 1309–1320, 2018.
- [118] W. Erin and W. Ye, “Blogging the brand: Meaning transfer and the case of weight watchers’ online community,” *Journal of Brand Management*, vol. 23, no. 4, pp. 457–471, 2016.
- [119] B. Liang, Y. Wang, and M.-H. Tsou, “A ”fitness” theme may mitigate regional prevalence of overweight and obesity: Evidence from google search and tweets,” *Journal of Health Communication*, vol. 24, no. 9, pp. 683–692, 2019.

- [120] A. Sarker, G. Gonzalez-Hernandez, Y. Ruan, and J. Perrone, “Machine learning and natural language processing for geolocation-centric monitoring and characterization of opioid-related social media chatter,” *JAMA Network Open*, vol. 2, no. 11, e1914672–e1914672, 2019.
- [121] Y. Fan, Y. Zhang, Y. Ye, X. Li, and W. Zheng, “Social media for opioid addiction epidemiology: Automatic detection of opioid addicts from twitter and case studies,” in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 1259–1267.
- [122] S. Chancellor, G. Nitzburg, A. Hu, F. Zampieri, and M. De Choudhury, “Discovering alternative treatments for opioid use recovery using social media,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–15.
- [123] M. Chary, N. Genes, C. Giraud-Carrier, C. Hanson, L. S. Nelson, and A. F. Manini, “Epidemiology from tweets: Estimating misuse of prescription opioids in the usa from social media,” *Journal of Medical Toxicology*, vol. 13, no. 4, pp. 278–286, 2017.
- [124] A. Lamb, M. J. Paul, and M. Dredze, “Investigating twitter as a source for studying behavioral responses to epidemics.,” in *AAAI Fall Symposium: Information Retrieval and Knowledge Discovery in Biomedical Text*, Citeseer, 2012, pp. 81–83.
- [125] R. Schwartz, M. Sap, I. Konstas, L. Zilles, Y. Choi, and N. A. Smith, “The effect of different writing tasks on linguistic style: A case study of the roc story cloze task,” *arXiv preprint arXiv:1702.01841*, 2017.

- [126] S. Pandrekar, X. Chen, G. Gopalkrishna, A. Srivastava, M. Saltz, J. Saltz, and F. Wang, “Social media based analysis of opioid epidemic using reddit,” in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2018, 2018, p. 867.
- [127] M. E. McCombs and D. L. Shaw, “The agenda-setting function of mass media,” *Public Opinion Quarterly*, vol. 36, no. 2, pp. 176–187, 1972.
- [128] E. E. McGinty, A. Kennedy-Hendricks, J. Baller, J. Niederdeppe, S. Gollust, and C. L. Barry, “Criminal activity or treatable health condition? news media framing of opioid analgesic abuse in the united states, 1998-2012,” *Psychiatric Services*, vol. 67, no. 4, pp. 405–411, 2016.
- [129] A. Kennedy-Hendricks, J. Levin, E. Stone, E. E. McGinty, S. E. Gollust, and C. L. Barry, “News media reporting on medication treatment for opioid use disorder amid the opioid epidemic,” *Health Affairs*, vol. 38, no. 4, pp. 643–651, 2019.
- [130] S. Nygaard, “Boundary work: Intermedia agenda-setting between right-wing alternative media and professional journalism,” *Journalism Studies*, vol. 21, no. 6, pp. 766–782, 2020.
- [131] E. Willis and C. Painter, “Conceptualization of the public health model of reporting through application: The case of the cincinnati enquirer’s heroin beat,” *Health Communication*, pp. 1–10, 2020.
- [132] —, “The needle and the damage done: Framing the heroin epidemic in the cincinnati enquirer,” *Health Communication*, vol. 34, pp. 661–671, 2019.

- [133] E. Riloff and J. Wiebe, “Learning extraction patterns for subjective expressions,” in *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 2003, pp. 105–112.
- [134] J. Wiebe, T. Wilson, and C. Cardie, “Annotating expressions of opinions and emotions in language,” *Language Resources and Evaluation*, vol. 39, no. 2, pp. 165–210, 2005.
- [135] W. He, L. J. Larsen, *et al.*, *Older Americans with a disability, 2008-2012*. US Census Bureau Washington, DC, 2014.
- [136] J. F. Kraus, M. A. Black, N. Hessol, P. Ley, W. Rokaw, C. Sullivan, S. Bowers, S. KNOWLTON, and A. MARSHALL, “The incidence of acute brain injury and serious impairment in a defined population,” *American Journal of Epidemiology*, vol. 119, no. 2, pp. 186–201, 1984.
- [137] K. M. Langa, M. E. Chernew, M. U. Kabeto, A. Regula Herzog, M. Beth Ofstedal, R. J. Willis, R. B. Wallace, L. M. Mucha, W. L. Straus, and A. M. Fendrick, “National estimates of the quantity and cost of informal caregiving for the elderly with dementia,” *Journal of General Internal Medicine*, vol. 16, no. 11, pp. 770–778, 2001.
- [138] M. J. Prince, A. Wimo, M. M. Guerchet, G. C. Ali, Y.-T. Wu, and M. Prina, “World alzheimer report 2015-the global impact of dementia: An analysis of prevalence, incidence, cost and trends,” 2015.

- [139] J. Chodosh, D. B. Petitti, M. Elliott, R. D. Hays, V. C. Crooks, D. B. Reuben, J. Galen Buckwalter, and N. Wenger, "Physician recognition of cognitive impairment: Evaluating the need for improvement," *Journal of the American Geriatrics Society*, vol. 52, no. 7, pp. 1051–1059, 2004.
- [140] A. Bradford, M. E. Kunik, P. Schulz, S. P. Williams, and H. Singh, "Missed and delayed diagnosis of dementia in primary care: Prevalence and contributing factors," *Alzheimer Disease and Associated Disorders*, vol. 23, no. 4, p. 306, 2009.
- [141] S. McPherson and G. Schoephoerster, "Screening for dementia in a primary care practice.," *Minnesota Medicine*, vol. 95, no. 1, pp. 36–40, 2012.
- [142] J. E. Galvin and C. H. Sadowsky, "Practical guidelines for the recognition and diagnosis of dementia," *The Journal of the American Board of Family Medicine*, vol. 25, no. 3, pp. 367–382, 2012.
- [143] M. Boustani, B. Peterson, L. Hanson, R. Harris, and K. N. Lohr, "Screening for dementia in primary care: A summary of the evidence for the us preventive services task force," *Annals of Internal Medicine*, vol. 138, no. 11, pp. 927–937, 2003.
- [144] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, and E. Kokmen, "Mild cognitive impairment: Clinical characterization and outcome," *Archives of Neurology*, vol. 56, no. 3, pp. 303–308, 1999.
- [145] S. Katz, A. B. Ford, R. W. Moskowitz, B. A. Jackson, and M. W. Jaffe, "Studies of illness in the aged: The index of adl: A standardized measure of biological and psychosocial function," *Journal of the American Medical Association*, vol. 185, no. 12, pp. 914–919, 1963.

- [146] M. P. Lawton and E. M. Brody, “Assessment of older people: Self-maintaining and instrumental activities of daily living,” *The Gerontologist*, vol. 9, no. 3_Part_1, pp. 179–186, 1969.
- [147] I. Hartigan, “A comparative review of the katz adl and the barthel index in assessing the activities of daily living of older people,” *International Journal of Older People Nursing*, vol. 2, no. 3, pp. 204–212, 2007.
- [148] M. Yang, X. Ding, and B. Dong, “The measurement of disability in the elderly: A systematic review of self-reported questionnaires,” *Journal of the American Medical Directors Association*, vol. 15, no. 2, 150–e1, 2014.
- [149] M. E. Mlinac and M. C. Feng, “Assessment of activities of daily living, self-care, and independence,” *Archives of Clinical Neuropsychology*, vol. 31, no. 6, pp. 506–516, 2016.
- [150] G. A. Marshall, R. E. Amariglio, R. A. Sperling, and D. M. Rentz, “Activities of daily living: Where do they fit in the diagnosis of alzheimer’s disease?” *Neurodegenerative Disease Management*, vol. 2, no. 5, pp. 483–491, 2012.
- [151] M. De Marco, L. Beltrachini, A. Biancardi, A. F. Frangi, and A. Venneri, “Machine-learning support to individual diagnosis of mild cognitive impairment using multimodal mri and cognitive assessments,” *Alzheimer Disease & Associated Disorders*, vol. 31, no. 4, pp. 278–286, 2017.
- [152] K.-H. Thung, P.-T. Yap, E. Adeli, S.-W. Lee, D. Shen, A. D. N. Initiative, *et al.*, “Conversion and time-to-conversion predictions of mild cognitive impairment using low-rank affinity pursuit denoising and matrix completion,” *Medical Image Analysis*, vol. 45, pp. 68–82, 2018.

- [153] S. H. Hojjati, A. Ebrahimzadeh, A. Khazaei, A. Babajani-Feremi, A. D. N. Initiative, *et al.*, “Predicting conversion from mci to ad using resting-state fmri, graph theoretical approach and svm,” *Journal of Neuroscience Methods*, vol. 282, pp. 69–80, 2017.
- [154] A. So, D. Hooshyar, K. W. Park, and H. S. Lim, “Early diagnosis of dementia from clinical data by machine learning techniques,” *Applied Sciences*, vol. 7, no. 7, p. 651, 2017.
- [155] S. Spasov, L. Passamonti, A. Duggento, P. Lio, N. Toschi, A. D. N. Initiative, *et al.*, “A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer’s disease,” *Neuroimage*, vol. 189, pp. 276–287, 2019.
- [156] M. Asgari, J. Kaye, and H. Dodge, “Predicting mild cognitive impairment from spontaneous spoken utterances,” *Alzheimer’s & Dementia: Translational Research & Clinical Interventions*, vol. 3, no. 2, pp. 219–228, 2017.
- [157] H. H. Alahmadi, Y. Shen, S. Fouad, C. D. B. Luft, P. Bentham, Z. Kourtzi, and P. Tino, “Classifying cognitive profiles using machine learning with privileged information in mild cognitive impairment,” *Frontiers in Computational Neuroscience*, vol. 10, p. 117, 2016.
- [158] P. Bhatkoti and M. Paul, “Early diagnosis of alzheimer’s disease: A multi-class deep learning framework with modified k-sparse autoencoder classification,” in *2016 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, IEEE, 2016, pp. 1–5.

- [159] T. Pereira, F. L. Ferreira, S. Cardoso, D. Silva, A. de Mendonça, M. Guerreiro, and S. C. Madeira, “Neuropsychological predictors of conversion from mild cognitive impairment to alzheimer’s disease: A feature selection ensemble combining stability and predictability,” *BMC Medical Informatics and Decision Making*, vol. 18, no. 1, pp. 1–20, 2018.
- [160] T. Pereira, L. Lemos, S. Cardoso, D. Silva, A. Rodrigues, I. Santana, A. de Mendonca, M. Guerreiro, and S. C. Madeira, “Predicting progression of mild cognitive impairment to dementia using neuropsychological data: A supervised learning approach using time windows,” *BMC Medical Informatics and Decision Making*, vol. 17, no. 1, pp. 1–15, 2017.
- [161] S. Kato, H. Endo, A. Homma, T. Sakuma, and K. Watanabe, “Early detection of cognitive impairment in the elderly based on bayesian mining using speech prosody and cerebral blood flow activation,” in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2013, pp. 5813–5816.
- [162] S. Amra, J. C. O’Horo, T. D. Singh, G. A. Wilson, R. Kashyap, R. Petersen, R. O. Roberts, J. D. Fryer, A. A. Rabinstein, and O. Gajic, “Derivation and validation of the automated search algorithms to identify cognitive impairment and dementia in electronic health records,” *Journal of Critical Care*, vol. 37, pp. 202–205, 2017.
- [163] H. Liu, S. J. Bielinski, S. Sohn, S. Murphy, K. B. Waghlikar, S. R. Jonnalagadda, K. Ravikumar, S. T. Wu, I. J. Kullo, and C. G. Chute, “An information extraction framework for cohort identification using electronic health records,” *AMIA Summits on Translational Science Proceedings*, vol. 2013, p. 149, 2013.

- [164] M. Torii, K. Waghlikar, and H. Liu, "Using machine learning for concept extraction on clinical documents from multiple data sources," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 580–587, 2011.
- [165] D. Neu, H. Kajosch, P. Peigneux, P. Verbanck, P. Linkowski, and O. Le Bon, "Cognitive impairment in fatigue and sleepiness associated conditions," *Psychiatry Research*, vol. 189, no. 1, pp. 128–134, 2011.
- [166] A. P. Spira, L. P. Chen-Edinboro, M. N. Wu, and K. Yaffe, "Impact of sleep on the risk of cognitive decline and dementia," *Current Opinion in Psychiatry*, vol. 27, no. 6, p. 478, 2014.
- [167] P. Jean-Pierre, "Management of cancer-related cognitive dysfunction-conceptualization challenges and implications for clinical research and practice," *US Oncology*, vol. 6, p. 9, 2010.
- [168] A. Magnuson, S. Mohile, and M. Janelins, "Cognition and cognitive impairment in older adults with cancer," *Current Geriatrics Reports*, vol. 5, no. 3, pp. 213–9, 2016.
- [169] D. Jones, E. Vichaya, X. Wang, M. Sailors, C. Cleeland, and J. Wefel, "Acute cognitive impairment in patients with multiple myeloma undergoing autologous hematopoietic stem cell transplant," *Cancer*, vol. 119, no. 23, pp. 4188–95, 2013.
- [170] A. Pakzad, N. Obad, H. Espedal, D. Stieber, O. Keunen, and P. Sakariassen, "Bevacizumab treatment for human glioblastoma. can it induce cognitive impairment?" *Neuro-Oncology*, vol. 16, no. 5, pp. 754–6, 2014.

- [171] Y. Cheng, Y. Jin, F. Unverzagt, L. Su, L. Yang, and F. Ma, “The relationship between cholesterol and cognitive function is homocysteine-dependent,” *Clinical Interventions in Aging*, vol. 9, pp. 1823–9, 2014.
- [172] L. Zheng, W. MacK, H. Chui, L. Heflin, D. Mungas, and B. Reed, “Coronary artery disease is associated with cognitive decline independent of changes on magnetic resonance imaging in cognitively normal elderly adults,” *Journal of the American Geriatrics Society*, vol. 60, no. 3, pp. 499–504, 2012.
- [173] C. Iadecola, K. Yaffe, J. Biller, L. Bratzke, F. Faraci, and P. Gorelick, “Impact of hypertension on cognitive function: A scientific statement from the american heart association,” *Hypertension*, vol. 68, no. 6, 2016.
- [174] C.-H. Chiang, M.-P. Wu, C.-H. Ho, S.-F. Weng, C.-C. Huang, and W.-T. Hsieh, “Lower urinary tract symptoms are associated with increased risk of dementia among the elderly: A nationwide study,” *BioMed Research International*, vol. 2015, pp. 1–7, 2015.
- [175] M. Fenske, “Urinary cortisol excretion: Is it really a predictor of incident cognitive impairment?” *Neurobiology of Aging*, vol. 28, no. 11, pp. 1791–2, 2007.
- [176] E. Trachtenberg, T. Mashiach, R. Ben Hayun, T. Tadmor, T. Fisher, J. Aharon-Peretz, and E. Dann, “Cognitive impairment in hodgkin lymphoma survivors,” *British Journal of Haematology*, vol. 182, no. 5, pp. 670–8, 2018.
- [177] S. Park, S. Kim, J. Sung, K. Lee, K. Park, and S. Kim, “Nocturnal hypoxia in als is related to cognitive dysfunction and can occur as clusters of desaturations,” *PLoS One*, vol. 8, no. 9, pp. 1–5, 2013.

- [178] K. Kreiter, D. Copeland, G. Bernardini, J. Bates, S. Peery, and J. Claassen, “Predictors of cognitive dysfunction after subarachnoid hemorrhage,” *Stroke*, vol. 33, no. 1, pp. 200–8, 2002.
- [179] Y. Wang, J. Zhang, W. Hu, J. Li, J. Zhou, and J. Zhang, “The characteristics of cognitive impairment in subjective chronic tinnitus,” *Brain and Behavior*, vol. 8, no. 3, pp. 1–9, 2018.
- [180] C. Ma, Z. Yin, P. Zhu, J. Luo, X. Shi, and X. Gao, “Blood cholesterol in late-life and cognitive decline: A longitudinal study of the chinese elderly,” *Molecular Neurodegeneration*, vol. 12, no. 1, pp. 1–9, 2017.
- [181] C. Restrepo, S. Patel, V. Rethnam, E. Werden, J. Ramchand, and L. Churilov, “Left ventricular hypertrophy and cognitive function: A systematic review,” *Journal of Human Hypertension*, vol. 32, no. 3, pp. 171–9, 2018.
- [182] S. Udompanich, G. Lip, S. Apostolakis, and D. Lane, “Atrial fibrillation as a risk factor for cognitive impairment: A semi-systematic review,” *QJM: An International Journal of Medicine*, vol. 106, no. 9, pp. 795–802, 2013.
- [183] C. Tseng, W. Huang, C. Muo, and C. Kao, “Increased risk of dementia among chronic osteomyelitis patients,” *European Journal of Clinical Microbiology Infectious Diseases*, vol. 34, no. 1, pp. 153–9, 2014.
- [184] H. Stradecki-Cohan, C. Cohan, A. Raval, K. Dave, D. Reginensi, and R. Gittens, “Cognitive deficits after cerebral ischemia and underlying dysfunctional plasticity: Potential targets for recovery of cognition,” *Journal of Alzheimers Diseases*, vol. 60, no. s1, 2017.

- [185] J. Dodd, "Lung disease as a determinant of cognitive decline and dementia," *Alzheimer's Research Therapy*, vol. 7, no. 1, pp. 1–8, 2015.
- [186] V. Vasudevan, "Effectiveness of media and enforcement campaigns in increasing seat belt usage rates in a state with a secondary seat belt law," *Traffic Injury Prevention*, vol. 10, no. 4, pp. 330–339, 2009.
- [187] M. Leurer, "Lessons in media advocacy: A look back at saskatchewan's nursing education debate," *Policy, Politics Nursing Practice*, vol. 14, no. 2, pp. 86–96, 2013.
- [188] A. Gardner, "Clinic consortia media advocacy capacity: Partnering with the media and increasing policymaker awareness," *Journal of Health Communication*, vol. 15, no. 3, pp. 293–306, 2010.
- [189] L. Bou-Karroum, "Using media to impact health policy-making: An integrative systematic review," *Implementation Science*, vol. 12, no. 1, pp. 52–62, 2017.
- [190] M. Wakefield, B. Loken, and R. Hornik, "Use of mass media campaigns to change health behaviour," *The Lancet*, vol. 376, no. 9748, pp. 1261–1271, 2010.
- [191] H. Weishaar, "Why media representations of corporations matter for public health policy: A scoping review," *BMC Public Health*, vol. 16, no. 1, pp. 899–905, 2016.
- [192] N. Glenn, C. Champion, and J. Spence, "Qualitative content analysis of online news media coverage of weight loss surgery and related reader comments," *Clinical Obesity*, vol. 2, no. 5-6, pp. 125–131, 2012.
- [193] C. Patterson, "Content analysis of uk newspaper and online news representations of women's and men's binge drinking: A challenge for communicating

- evidence-based messages about single-episodic drinking?” *BMJ Open*, vol. 6, no. 12, pp. 13–24, 2016.
- [194] B. Liu and L. Zhang, “A survey of opinion mining and sentiment analysis,” in *Mining Text Data*, C. Aggarwal and C.-X. Zhai, Eds., New York: Springer, 2012, pp. 415–463.
- [195] *National library of medicine. medline: Description of the database*, Retrieved from, 2019. [Online]. Available: <https://www.nlm.nih.gov/bsd/medline.html>.
- [196] C. Lipscomb, “Medical subject headings (mesh),” *Bulletin of the Medical Library Association*, vol. 88, no. 3, pp. 265–276, 2000.
- [197] P. Schuyler, “The umls metathesaurus: Representing different views of biomedical concepts,” *Bulletin of the Medical Library Association*, vol. 81, no. 2, pp. 217–222, 1993.
- [198] O. Bodenreider, “The unified medical language system (umls): Integrating biomedical terminology,” *Nucleic Acids Research*, vol. 32, no. suppl₁, pp. 267–270, 2004.
- [199] M. Huang, “Public opinions toward diseases: Infodemiological study on news media data,” *Journal of Medical Internet Research*, vol. 20, no. 5, pp. 100–147, 2018.
- [200] Reuters, *Reuters site archive, united states*, Retrieved from: 2018. [Online]. Available: <https://www.Reuters.com>.
- [201] X. Chen, “Datamed—an open source discovery index for finding bio- medical datasets,” *Journal of the American Medical Informatics Association*, vol. 25, no. 3, pp. 300–308, 2018.

- [202] W. Hahn, “The effect of media attention on concern for and medical management of methicillin-resistant staphylococcus aureus: A multime- thod study,” *Journal of Public Health Management and Practice*, vol. 15, no. 2, pp. 150–159, 2009.
- [203] R. Mahabir, “News coverage, digital activism, and geographical saliency: A case study of refugee camps and volunteered geographical infor- mation,” *PLoS One*, vol. 13, no. 11, pp. 206–225, 2018.
- [204] C. Buckton, “A discourse network analysis of uk newspaper cover- age of the ”sugar tax” debate before and after the announcement of the soft drinks industry levy,” *BMC Public Health*, vol. 19, no. 1, pp. 490–501, 2019.
- [205] M. Motta, T. Callaghan, and S. Sylvester, “Knowing less but presuming more: Dunning-kruger effects and the endorsement of anti-vaccine policy attitudes,” *Social Science Medicine*, vol. 211, pp. 274–281, 2018.
- [206] M. Mantyla, D. Graziotin, and M. Kuutilla, “The evolution of sentiment analysis-a review of research topics, venues, and top cited papers,” *Computer Science Re- view*, vol. 27, pp. 16–32, 2018.
- [207] M. Rastegar-Mojarad, “Detecting signals in noisy data-can ensemble classifiers help identify adverse drug reaction in tweets,” in *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*, 2016.
- [208] A. Hopper and M. Uriyo, “Using sentiment analysis to review patient satisfac- tion data located on the internet,” *Journal of Health Organization and Management*, vol. 29, no. 2, pp. 221–233, 2015.
- [209] E. Clark, *A sentiment analysis of breast cancer treatment experiences and health- care perceptions across twitter*, arXiv preprint arXiv:1805.09959; 2018.

- [210] D. Trilling, *Doing computational social science with python: An introduction*, Available at SSRN 2737682; 2018.
- [211] C. Hutto and E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.
- [212] M. Huang, “Technological innovations in disease management: Text mining us patent data from 1995 to 2017,” *Journal of Medical Internet Research*, vol. 21, no. 4, pp. 133–136, 2019.
- [213] K. He, “Understanding the patient perspective of epilepsy treatment through text mining of online patient support groups,” *Epilepsy Behavior*, vol. 94, pp. 65–71, 2019.
- [214] D. Blei, A. Ng, and M. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, no. January, pp. 993–1022, 2003.
- [215] Y. Kim, “Clinical progress of human papillomavirus genotypes and their persistent infection in subjects with atypical squamous cells of undetermined significance cytology: Statistical and latent dirichlet allocation analysis,” *Experimental and Therapeutic Medicine*, vol. 13, no. 6, pp. 3032–3038, 2017.
- [216] K. Stevens, “Exploring topic coherence over many models and many topics,” in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, 2012.

- [217] J. Schneider and M. Vlachos, “Topic modeling based on keywords and context,” in *Proceedings of the 2018 SIAM International Conference on Data Mining. SIAM*, 2018.
- [218] J. S. M. Package, Retrieved from, 2018. [Online]. Available: <https://github.com/JohnTailor/tkm>.
- [219] H. Liu, “Biolemmatizer: A lemmatization tool for morphological processing of biomedical text,” *Journal of Biomedical Semantics*, vol. 3, no. 1, pp. 3–11, 2012.
- [220] N. Suaysom and W. Gu, “Expert opinion and coherence based topic modeling,” *International Journal on Natural Language Computing (IJNLC) Vol*, vol. 7, 2018.
- [221] K. Kelland, *Study details how high fiber diets make for healthier lives*, Retrieved from, 2019. [Online]. Available: <https://www.reuters.com/article/us-health-fibre/study>.
- [222] L. Papaport, *Aerobic exercise eases depression, even in chronically ill*, Retrieved from, 2019. [Online]. Available: <https://www.reuters.com/article/us-health-depression>.
- [223] J. E. Ahlskog, Y. E. Geda, N. R. Graff-Radford, and R. C. Petersen, “Physical exercise as a preventive or disease-modifying treatment of dementia and brain aging,” in *Mayo Clinic Proceedings*, Elsevier, vol. 86, 2011, pp. 876–884.
- [224] C. Canning, “Exercise for falls prevention in parkinson disease: A randomized controlled trial,” *Neurology*, vol. 84, no. 3, pp. 304–312, 2015.
- [225] J. E. Ahlskog, “Aerobic exercise: Evidence for a direct brain effect to slow parkinson disease progression,” in *Mayo Clinic Proceedings*, Elsevier, vol. 93, 2018, pp. 360–372.

- [226] M. Maraki, “Mediterranean diet adherence is related to reduced probability of prodromal parkinson’s disease,” *Movement Disorders*, vol. 34, no. 1, pp. 48–57, 2019.
- [227] S. Gielen, “Exercise training in patients with heart disease: Review of beneficial effects and clinical recommendations,” *Progress in Cardiovascular Diseases*, vol. 57, no. 4, pp. 347–355, 2015.
- [228] C. Fiuza-Luces, “Exercise benefits in cardiovascular disease: Beyond attenuation of traditional risk factors,” *Nature Reviews Cardiology*, vol. 15, no. 12, pp. 731–743, 2018.
- [229] R. Estruch, “Primary prevention of cardiovascular disease with a mediterranean diet supplemented with extra-virgin olive oil or nuts,” *New England Journal of Medicine*, vol. 378, no. 25, pp. 34–46, 2018.
- [230] J. Hendry, *Drinking alcohol may keep leg arteries healthy*, Retrieved from, 2007. [Online]. Available: <https://www.reuters.com/article/us-alcohol-leg-arteries/drinking>.
- [231] K. Kelland, *Moderate drinkers have a better health, study finds*, Retrieved from, 2010. [Online]. Available: <https://www.reuters.com/article/us-alcohol/moderate>.
- [232] E. Mostofsky, “Key findings on alcohol consumption and a variety of health outcomes from the nurses’ health study,” *American Journal of Public Health*, vol. 106, no. 9, pp. 1586–1591, 2016.
- [233] K. Nebehay, *Air pollution a leading cause of cancer - u.n. agency*, Retrieved from, 2013. [Online]. Available: <https://www.reuters.com/article/us-cancer-pollution/air>.

- [234] K. Russell, *Poor mental health harming productivity, says oecd*, Retrieved from, 2011. [Online]. Available: <https://www.reuters.com/article/us-mental-work/poor-mental>.
- [235] L. Carroll, *Sexual harassment, abuse tied to real health effects*, Retrieved from, 2018. [Online]. Available: <https://www.reuters.com/article/us-health-assault-harassment/sexual>.
- [236] C. Crist, *Kids with asthma often leave doctor's office with unanswered questions*, Retrieved from, 2019. [Online]. Available: <https://www.reuters.com/article/us-health>.
- [237] L. Rapaport, *Pot-smoking on the rise among u.s. pregnant women*, Retrieved from, 2017. [Online]. Available: <https://www.reuters.com/article/us-health-pregnanc>.
- [238] P. Gilbert, *Depression: The Evolution of Powerlessness*. New York: Routledge, 2016.
- [239] U. D. of Health and H. Services, "Let's make the next generation tobacco-free: Your guide to the 50th anniversary surgeon general's report on smoking and health. atlanta: Us department of health and human services, centers for disease control and prevention," *National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health*, 2014.
- [240] R. Arrazola, "Tobacco use among middle and high school students-united states," *Morbidity and Mortality Weekly Report*, vol. 63, no. 45, pp. 10–21, 2013.
- [241] C. for Disease Control, *Preventing tobacco use among youth and young adults*. Centers for Disease Control and Prevention, 2012.

- [242] C. Weinstock, *Cutting back on alcohol can prevent cancers: Experts*, Retrieved from: 2017. [Online]. Available: <https://www.reuters.com/article/us-health-cancer-alcohol/>.
- [243] A. Harding, *Moderate drinking helps heart, but don't binge*, Retrieved from: 2010. [Online]. Available: <https://www.reuters.com/article/us-moderate-drinking/moderate>.
- [244] J. Hojsted and P. Sjogren, "Addiction to opioids in chronic pain patients: A literature review," *European Journal of Pain*, vol. 11, no. 5, pp. 490–518, 2007.
- [245] G. Rassool, *Alcohol and Drug Misuse: A Guide for Health and Social Care Professionals*. New York: Routledge, 2017.
- [246] A. Fan, "Drinking and driving among adults in the united states: Results from the 2012-2013 national epidemiologic survey on alcohol and related conditions-iii," *Accident Analysis Prevention*, vol. 125, pp. 49–55, 2019.
- [247] C. Patten, "Addressing community health needs through community engagement research advisory boards," *Journal of Clinical and Translational Science*, vol. 3, no. s1, pp. 82–82, 2019.
- [248] J. Khubchandani, "Community-engaged strategies to increase diversity of participants in health education research," *Health Promotion Practice*, vol. 17, no. 3, pp. 323–327, 2016.
- [249] J. Balls-Berry, "Using garden cafes to engage community stakeholders in health research," *PLoS One*, vol. 13, no. 8, pp. 2–13, 2018.

VITA

Saria Goudarzvandis from north of Iran, Gilan. She graduated from Islamic Azad University of Tehran North Tehran Branch, Iran in 2014. In January 2017, she moved to the USA to start her Interdisciplinary Ph.D. in Computer Science from the School of Computing and Engineering, University of Missouri-Kansas City, under the supervision of Dr. Yugyung Lee. Saria Goudarzvand's research interests include Machine Learning, Knowledge Discovery, Natural Language Processing. She is currently leading two projects with her doctoral advisor, namely "Implicit Bias in STEM" and "Understanding Greek Literature using Deep Learning." Saria has received multiple awards, such as ranked first place, and Audience choice award in Healthcare Innovation at Blue Cross and Blue Shield Of Kansas City and SGS Research Award. She did research internship in Mayo Clinic on the area of Natural Language Processing. She has published in multiple top conference and Journals such as AAI, Applied Intelligence, BMC medical informatics and decision making. She is working in the area of Artificial Intelligence and Natural Language Processing as Machine Learning Scientist at Expedia.