BRIDGING AI AND PERSONALIZATION: FROM SOCIAL MEDIA INSIGHTS TO TARGETED MARKETING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

YUSUF MÜCAHIT ÇETINKAYA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

JANUARY 2025

Approval of the thesis:

BRIDGING AI AND PERSONALIZATION: FROM SOCIAL MEDIA INSIGHTS TO TARGETED MARKETING

submitted by **YUSUF MÜCAHIT ÇETINKAYA** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Naci Emre Altun
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering** _____

Prof. Dr. İsmail Hakkı Toroslu
Supervisor, **Computer Engineering** _____

Examining Committee Members:

Prof. Dr. Ahmet Coşar
Computer Engineering, Ankara Medipol University _____

Prof. Dr. İsmail Hakkı Toroslu
Computer Engineering, METU _____

Assoc. Prof. Dr. Hande Alemdar
Computer Engineering, METU _____

Assist. Prof. Dr. Engin Demir
Computer Engineering, Hacettepe University _____

Assist. Prof. Dr. Çağrı Toraman
Computer Engineering, METU _____

Date:10.01.2025

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Yusuf Mücahit Çetinkaya

Signature :

ABSTRACT

BRIDGING AI AND PERSONALIZATION: FROM SOCIAL MEDIA INSIGHTS TO TARGETED MARKETING

Çetinkaya, Yusuf Mücahit
Ph.D., Department of Computer Engineering
Supervisor: Prof. Dr. İsmail Hakkı Toroslu

January 2025, 123 pages

In today's digital landscape, integrating artificial intelligence (AI) and personalization is vital for understanding and engaging audiences on social media platforms. This thesis investigates scalable AI-driven methods to extract meaningful insights from social media content and transform them into actionable strategies for targeted marketing. The research begins with Twitter account classification, where AI models leverage metadata to differentiate individual and organizational accounts. Building upon this, the thesis introduces a framework for personalized paragraph generation aimed at creating compelling landing pages. Using advanced text generation techniques, AI systems produce coherent, relevant, and engaging content tailored to user needs. The study further explores targeted marketing strategies by applying text analysis methods to design and generate personalized landing pages informed by social media insights. This approach bridges the gap between audience understanding and practical marketing applications. To enable deeper content personalization, the thesis develops techniques for aspect-based sentiment analysis. These methods facilitate scalable and detailed sentiment evaluation, addressing challenges such as echo chambers and bias while improving content relevance and fairness. Additionally, the thesis

proposes a programmable, stance-directed AI architecture that generates human-like, personalized social media content. This framework aligns AI outputs with user preferences and stances, fostering humanized and context-aware communication strategies. Through theoretical advancements and experimental validation, this work demonstrates how AI can bridge the gap between social media insights, user segmentation, and targeted marketing.

Keywords: AI-driven personalization, social media, targeted marketing, user segmentation

ÖZ

YAPAY ZEKA VE KİŞİSELLEŞTİRME: SOSYAL MEDYA İÇGÖRÜLERİNDEN HEDEFLİ PAZARLAMAYA

Çetinkaya, Yusuf Mücahit
Doktora, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. İsmail Hakkı Toroslu

Ocak 2025 , 123 sayfa

Günümüz dijital ortamında yapay zeka (AI) ve kişiselleştirmenin entegrasyonu, sosyal medya platformlarında kitleleri anlamak ve onlarla etkileşim kurmak için hayati öneme sahiptir. Bu tez, sosyal medya içeriğinden anlamlı içgörüler çıkarmak ve bunları hedeflenmiş pazarlama için uygulanabilir stratejilere dönüştürmek amacıyla ölçeklenebilir, yapay zeka destekli yöntemleri araştırmaktadır. Araştırma, Twitter hesap sınıflandırması ile başlamaktadır; burada yapay zeka modelleri, bireysel ve kurumsal hesapları ayırt etmek için meta verilerden yararlanmaktadır. Bunun üzerine inşa edilen tez, etkileyici açılış sayfaları oluşturmak için kişiselleştirilmiş paragraf üretim çerçevesi sunmaktadır. Gelişmiş metin üretim tekniklerini kullanarak, yapay zeka sistemleri kullanıcı ihtiyaçlarına yönelik tutarlı, ilgili ve ilgi çekici içerikler üretmektedir. Çalışma, ayrıca, sosyal medya içgörülerinden faydalanarak kişiselleştirilmiş açılış sayfaları tasarlamak ve üretmek amacıyla metin analiz yöntemlerini uygulayarak hedeflenmiş pazarlama stratejilerini araştırmaktadır. Bu yaklaşım, kitleleri anlama ile pratik pazarlama uygulamaları arasındaki boşluğu kapatmaktadır. Daha detaylı içerik kişiselleştirmesini sağlamak için tez, konu bazlı duygu analizi tekniklerini geliştir-

mektedir. Bu yöntemler, yankı odaları ve önyargı gibi zorlukları ele alırken içerik alakalılığını ve adaletini iyileştirerek ölçeklenebilir ve detaylı duygu değerlendirmesini mümkün kılar. Ek olarak, bu çalışma, insan benzeri, kişiselleştirilmiş sosyal medya içeriği üreten programlanabilir, duruş-yönelimli bir yapay zeka mimarisi önermektedir. Bu çerçeve, yapay zeka çıktılarının kullanıcı tercihleri ve duruşları ile uyumlu olmasını sağlayarak insanlaştırılmış ve bağlama duyarlı iletişim stratejilerini teşvik eder. Teorik ilerlemeler ve deneysel doğrulama yoluyla bu çalışma, yapay zekanın sosyal medya içgörüleri, kullanıcı segmentasyonu ve hedeflenmiş pazarlama arasındaki boşluğu nasıl kapatabileceğini göstermektedir.

Anahtar Kelimeler: yapay zeka destekli kişiselleştirme, sosyal medya, hedeflenmiş pazarlama, kullanıcı segmentasyonu

To my wife, son, and whole family members...

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Dr. İsmail Hakkı Toroslu, for his invaluable guidance, support, and encouragement throughout my doctoral journey. His mentorship has been instrumental in shaping my research and academic growth.

I would also like to extend my sincere thanks to Prof. Dr. Hasan Davulcu for his collaboration and hospitality during my time at Arizona State University (ASU). I am profoundly grateful for TUBITAK-2214/A support of TÜBİTAK for my visit to ASU.

I deeply appreciate the invaluable comments, feedback, and insights provided by my committee members. Their thorough reviews and constructive critiques significantly enhanced the quality and rigor of my research.

On a personal note, I am forever thankful to my wife and son for their endless patience, love, and unwavering support throughout this challenging journey. My heartfelt gratitude also goes to my parents and brother, whose constant encouragement and belief in me have always been a source of strength.

Finally, I extend my thanks to all my colleagues, friends, and anyone who has contributed to my work, directly or indirectly, over the years.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xv

LIST OF FIGURES xviii

LIST OF ABBREVIATIONS xx

CHAPTERS

1 INTRODUCTION 1

1.1 Motivation and Problem Definition 2

1.2 Contributions and Novelties 4

1.3 The Outline of the Thesis 4

2 LITERATURE SURVEY 7

2.1 Twitter Account Classification 7

2.2 Targeted Marketing on Social Media 11

2.3 Aspect-Based Sentiment Analysis 15

2.4 AI-Based Content Generation 17

3 TWITTER ACCOUNT CLASSIFICATION USING ACCOUNT METADATA: ORGANIZATION VS. INDIVIDUAL 23

3.1 Method 25

3.1.1 Data Collection 25

3.1.2 Data Analysis 27

3.1.3 Data Preprocessing 32

3.1.4 Classification 34

3.2 Experiments and Evaluation 36

3.3 Discussion 40

4 TARGETED MARKETING ON SOCIAL MEDIA: UTILIZING TEXT ANALYSIS TO CREATE PERSONALIZED LANDING PAGES 43

4.1 Data and Methodology 47

4.1.1 Data collection 48

4.1.2 Classifying Related Tweets 51

4.1.3 Topic modeling for user's priorities 55

4.1.4 Generating a coherent landing page paragraph 56

4.2 Experiments 57

4.2.1 Related Tweet Classification Experiments 57

4.2.2 User Priority Extraction Experiments 60

4.2.3 Landing Page Paragraph Generation Experiments 61

4.2.4 Human evaluation experiments 62

4.3 Employing Large Language Models for Personalized Landing Pages . 63

4.4 Discussion 68

5 MASKING THE BIAS: FROM ECHO CHAMBERS TO LARGE SCALE ASPECT-BASED SENTIMENT ANALYSIS 71

5.1 Methodology 73

5.1.1 Masked Aspect Sentiment Classification (MASC) 73

5.1.2 Weakly Labeled Dataset 75

5.1.2.1 U.S. Race Relations Dataset 76

5.1.2.2 Camp Labeling of Users 76

5.1.2.3 Bias and Sentiment Coding 78

5.1.2.4 Preparing U.S. Race Relations and Politics Datasets . . 79

5.2 Experiments and Evaluations 79

5.2.1 SemEval Datasets 79

5.2.2 U.S. Race Relations and Politics Datasets 80

5.2.3 Bias Test 81

5.2.4 New Terms Test 82

5.3 Discussion 83

6 TOWARDS A PROGRAMMABLE HUMANIZING AI THROUGH SCALABLE STANCE-DIRECTED ARCHITECTURE 85

6.1 Dataset and Camp Labeling 87

6.1.1 Camp Labeling of Users 87

6.1.2 Bias, Sentiment and Common Good Coding 88

6.1.3 Weakly Labeled Dataset 90

6.2 System Architecture 93

6.2.1 Stance-Directed Humanized LLM 93

6.2.2 Aspect-Based Sentiment Analysis 94

6.2.3 Toxic Content Detection 94
6.3 Experimental Evaluation 95
6.3.1 Comparison with pre-trained LLMs 97
6.4 Discussion 98
7 CONCLUSION 101
REFERENCES 105
CURRICULUM VITAE 121

LIST OF TABLES

TABLES

Table 3.1 Humanizr and Demographer dataset statistics. 24

Table 3.2 Twint provides 18 different attributes for user profiles. The attributes used in this study are marked in bold. 26

Table 3.3 The statistics for the word lengths in textual features in Demographer dataset. 31

Table 3.4 The statistics of the numerical features in Demographer dataset. . . . 31

Table 3.5 Sentencepiece tokenization results for some examples. 34

Table 3.6 Performance measurements of each model under k-fold cross-validation. 37

Table 3.7 Performance details of the proposed model under k-fold cross-validation. 38

Table 3.8 Comparison of the proposed model using numeric, textual, and textual+numeric features on Demographer full dataset under k-fold cross-validation. 38

Table 4.1 The tabulated count of the posts procured from Facebook groups and pages concerning the diabetes. 48

Table 4.2 Some of the keywords extracted by YAKE! from collected Facebook posts. Is keyword column is the label provided by domain experts. 49

Table 4.3 Some of the entries in the codebook created by domain experts of the sample use case. 50

Table 4.4 Codebook summary. 51

Table 4.5 Dataset summary which is labeled using K-Means clustering on top of BERT, RoBERTa embedding and corresponding sentence-transformer models displaying related/unrelated clusters. 53

Table 4.6 Cosine similarities between 'sweet' words' BERT embeddings in different contexts. 53

Table 4.7 The codebook-derived marketable aspects of the product and the accompanying general sentences for each aspect. 56

Table 4.8 Generated paragraph for the tweet *"Skyrocketing drug prices are hurting patients and costing taxpayers. Millions of Americans don't take their medications as prescribed due to the price."* using two different approaches. 58

Table 4.9 Relatedness classification performance results of the models with different threshold values. 59

Table 4.10 The ranked scores of the top-five topics predicted for a given tweet based on the user-provided ranks, which are sorted by their rank-oriented scores with different tolerance values. 60

Table 4.11 The identified topics by GuidedLDA are exhibited using given seeds, displaying the top 7 n-grams. 61

Table 4.12 The divergence and cosine similarity metrics of the selected next-sentence embedding model with respect to the average of the previous sentence. 62

Table 4.13 The results of the human evaluation that compares the effectiveness of the static landing page and the personalized landing page proposed in this study. 63

Table 5.1 ATSC accuracy results (%) on SemEval datasets compared with MaskedABSA models. 80

Table 5.2 Cross-domain ATSC results (%) on SemEval datasets comparing with the top model in the literature. 80

Table 5.3 Race and Politics datasets distribution. 80

Table 5.4 Cross-domain test results on the U.S. Race Relations and Politics datasets using MaskedABSA. 81

Table 5.5 Accuracy with replacements of new terms. 82

Table 6.1 Performance of the classifier on toxicity detection 96

Table 6.2 Comparable reduction in toxicity on sensitive domain 96

LIST OF FIGURES

FIGURES

Figure 3.1 Word cloud of Demographer dataset, containing top 250 occurred words in individual (left) and organization (right) accounts. . . . 26

Figure 3.2 The distribution of individual and organization accounts in the Demographer dataset comparing the follower count vs. (a) followee, (b) tweet, and (c) like counts. Note that the first chart in each row is for the whole collection of individual and organizational accounts. 28

Figure 3.3 The distribution of individual and organization accounts in the Demographer dataset comparing tweet count vs. (a) like, (b) media, and (c) following counts. Note that the first chart in each row is for the whole collection of individual and organization accounts. 29

Figure 3.4 Analysis on accounts for existence of URL, location and bio attributes in Demographer dataset 32

Figure 3.5 Original Byte-Pair Compression Algorithm 33

Figure 3.6 Multi-layer perceptron (MLP) for numeric features and long short-term memory (LSTM) for textual features are used. 35

Figure 4.1 The extended version of the online marketing activity funnel. . . 45

Figure 4.2 The visualization of the PCA applied to the tweets' BERT embeddings in 150 clusters, viewed from different angles. Best viewed in color. 52

Figure 4.3 The silhouette scores for clustering all collected tweets utilizing the K-Means algorithm while varying the number of clusters k. 52

Figure 4.4 Pre-trained language model based classifiers' model diagram. . . 54

Figure 4.5 Hallucination Example - Proceed from Tweet Source 65

Figure 4.6 Hallucination Example - Proceed From Knowledge Base 65

Figure 4.7 Template Description in the Prompt 66

Figure 4.8 Over-Specificity Example - Proceed From Tweet Source 67

Figure 5.1 Example of aspect-based sentiment annotation. 74

Figure 5.2 Visualization of polarized camps in U.S. Race Relations 77

Figure 5.3 Only Barrier-crossing users' network from U.S. Race Relations dataset 77

Figure 5.4 An example of a stance annotated sentence from the BLM camp. 79

Figure 6.1 Retweet Network Graph of Seven Major Communities, 17 Sub Communities, and Barrier-crossing Users in U.S. Race Relations network labeled with corresponding sub-community IDs. 86

Figure 6.2 Word cloud of the codebook displaying (left-to-right) positive/negative phrases, common good/public interest and othering/divisive phrases. Sizes represent their frequencies. 89

Figure 6.3 Weakly labeled dataset preparation rules based on the codebook for the language models with running examples. 90

Figure 6.4 The process of generating a non-toxic message thru a stance-directed prompt on certain entities and issues. 92

LIST OF ABBREVIATIONS

AAAI	Association for the Advancement of Artificial Intelligence
ABSA	Aspect-Based Sentiment Analysis
ACM	Association for Computing Machinery
ACOS	Aspect-Category-Opinion-Sentiment
AI	Artificial Intelligence
AIGC	AI-Generated Content
ALM	All Lives Matter
ALSC	Aspect-Level Sentiment Classification
API	Application Programming Interface
APTGN	Ad-Profile-based Title Generation Network
ASPE	Aspect Sentiment Pair Extraction
ASTE	Aspect Sentiment Triplet Extraction
ATE	Aspect Term Extraction
ATSC	Aspect Term Sentiment Classification
BART	Bidirectional and Auto-Regressive Transformers
BERT	Bidirectional Encoder Representations from Transformers
BLM	Black Lives Matter
CBOW	Continuous Bag-of-Words
CRF	Conditional Random Fields
DLGM	Disentangled Linguistic Graph Model
DTC-GT	Direct-to-Consumer Genetic Testing
GDPR	General Data Protection Regulation
GNN	Graph Neural Network
GPT	Generative Pre-trained Transformer

GPU	Graphics Processing Unit
IDF	Inverse Document Frequency
LDA	Latent Dirichlet Allocation
LLM	Large Language Model
LSTM	Long Short-Term Memory
MAB	Multi-Armed Bandit
MASC	Masked Aspect Sentiment Classification
MLP	Multi-Layer Perceptron
NER	Named Entity Recognition
NLG	Natural Language Generation
NLP	Natural Language Processing
OT	Opinion Terms
PCA	Principal Component Analysis
PHVM	Planning-based Hierarchical Variational Model
RLHF	Reinforcement Learning with Human Feedback
RoBERTa	A Robustly Optimized BERT Pretraining Approach
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
T5	Text-to-Text Transfer Transformer
TADW	Text-Associated DeepWalk
TUBITAK	The Scientific and Technological Research Council of Turkey
UMAP	Uniform Manifold Approximation and Projection
YAKE	Yet Another Keyword Extractor

CHAPTER 1

INTRODUCTION

Everything in the universe exists with its opposite—light and darkness, chaos and order, connection and isolation. The key to achieving harmony lies in understanding the differences and similarities between these opposing forces. For those who can master this understanding, the path to endless bliss unfolds. This principle takes on profound significance in the digital era, where social media serves as both a connector and a divider. The ever-changing dynamics of social media demand more than a singular viewpoint to comprehend its complexity or to leverage it effectively for targeted engagement.

Social media platforms have become a global stage for interaction, information exchange, and influence. However, these platforms also amplify polarization, echo chambers, and misinformation, creating a fragmented digital landscape. To navigate this complexity, it is essential to go beyond isolated approaches and consider a holistic orchestration of network structures, text contexts, audience priorities, and demographic factors. This integrated perspective is vital for unlocking the potential of personalization and delivering impactful, targeted advertisements. Businesses, researchers, and policymakers can create strategies that foster meaningful engagement and social harmony through this synthesis.

At its core, this research also strives to align AI systems with human values, creating tools that bridge gaps rather than widen them. By harmonizing opposites and addressing challenges such as bias, echo chambers, and misinformation, AI can evolve as a force for unity and understanding. The proposed framework humanizes AI, ensuring it not only performs efficiently but also resonates ethically with users, fostering trust and transparency. This alignment between technological advancement and human

needs is pivotal for creating systems that contribute positively to society.

Looking to the future, this thesis envisions AI as more than a tool for targeted marketing; it lays the groundwork for applications that extend into personalized education, adaptive healthcare, and public policy. This work sets a foundation for AI systems capable of more profound, more intuitive human interaction by addressing challenges such as data privacy, ethical AI design, and global scalability. In doing so, it anticipates a future where technology transforms complexity into coherence, creating tailored solutions and meaningful connections across a fragmented digital world.

This thesis explores the intricate interplay of these elements, presenting a comprehensive framework for harnessing AI to unlock the potential of personalization in social media. It begins with the critical task of Twitter account classification, employing metadata to distinguish individual users from organizations. From this foundation, the study delves into personalized content generation, addressing the needs and preferences of specific audiences through coherent and targeted landing page creation.

Furthermore, the work confronts challenges such as bias and echo chambers, introducing advanced methods for aspect-based sentiment analysis that ensure fairness and accuracy. At its core lies a programmable, stance-directed AI architecture that aligns content generation with user preferences and fosters human-like, context-aware communication. By addressing these challenges in an integrated and orchestrated manner, this research demonstrates how AI can bridge the gap between raw social media data and actionable, user-focused marketing strategies.

Through this symphony of insights—spanning networks, text, priorities, and demographics—this thesis reveals the immense potential of AI to transform social media engagement into a powerful tool for meaningful, ethical, and effective communication.

1.1 Motivation and Problem Definition

Social media has become a dominant platform for interaction, content sharing, and information dissemination. The massive growth of social media platforms has opened

opportunities for businesses, researchers, and organizations to understand their audiences better and deliver relevant, highly tailored content. However, effectively leveraging this wealth of data requires overcoming significant challenges related to user classification, bias, personalization, and content generation.

The motivation behind this research stems from the growing need to bridge AI and personalization for social media platforms. On the one hand, AI has shown great promise in analyzing large-scale data and generating content; on the other hand, personalization ensures that the results of these analyzes resonate with users' unique needs and preferences. By combining these two dimensions, businesses and organizations can deliver meaningful, coherent, targeted experiences to their audiences.

A key challenge lies in understanding who the users are. Social media platforms include diverse entities ranging from individuals to organizations, and distinguishing between these entities is critical for targeted communication strategies. Incorrect user classification leads to irrelevant content delivery and missed engagement opportunities.

Once the audience is accurately segmented, content coherence and personalization challenges arise. Personalized marketing strategies often rely on landing pages tailored to user preferences. However, generating such pages demands AI systems capable of producing coherent and relevant paragraphs that align with the interests of specific user groups.

Another critical problem addressed in this thesis is the bias and echo chamber effects inherent in social media content. Sentiment analysis methods often suffer from bias, leading to inaccurate user stances and preference assessments. This issue becomes more pronounced when analyzing large-scale aspect-based sentiment data, where ensuring fairness and accuracy is essential for effective personalization.

Finally, humanized and programmable AI systems that go beyond generic solutions are needed to deliver personalized, stance-aware content to social media users. Current AI solutions cannot often align content with user perspectives, limiting their effectiveness in fostering meaningful engagement.

1.2 Contributions and Novelties

To this end, this thesis tackles these challenges in a structured and integrated manner by addressing:

- User classification through metadata analysis to distinguish individual users from organizations,
- Targeted marketing through text analysis to create highly personalized landing pages,
- Bias mitigation and aspect-based sentiment analysis for fair and detailed content insights,
- Stance-directed, humanized AI for generating personalized content aligned with user needs and preferences.

By addressing these interconnected problems, this research aims to advance AI systems that integrate personalization, coherence, and fairness, ultimately bridging the gap between social media insights and targeted marketing solutions.

1.3 The Outline of the Thesis

This thesis is organized into seven chapters, each addressing a specific aspect of the research conducted. Chapter 2 provides a comprehensive review of the existing literature, focusing on the methodologies and advancements relevant to Twitter account classification, targeted marketing on social media, and aspect-based sentiment analysis.

Chapter 3 details the methods and experiments of classifying Twitter accounts using metadata to distinguish between individuals and organizations. This classification forms the foundation for effective user segmentation. Chapter 4 extends the discussion to targeted marketing by utilizing text analysis to design and generate personalized landing pages informed by social media insights. The methodologies for

creating coherent, relevant, and engaging paragraphs tailored to user preferences are thoroughly examined.

In Chapter 5, the thesis addresses the challenges of bias and echo chambers inherent in social media sentiment analysis. It introduces novel techniques for aspect-based sentiment analysis that enhance content relevance and fairness. Chapter 6 presents a programmable, stance-directed AI architecture that generates human-like, personalized content aligned with user preferences and stances, fostering meaningful engagement.

Finally, Chapter 7 concludes the thesis by summarizing this research's key findings and contributions. It discusses the implications for developing human-centered AI systems that prioritize personalization, coherence, and fairness and suggests directions for future research in the field.

CHAPTER 2

LITERATURE SURVEY

In this chapter, we present a thorough review of the existing research and methodologies that are closely aligned with this thesis's key themes. The primary aim of this survey is to uncover research gaps, showcase recent advancements, and contextualize our work within the broader academic discourse.

The chapter is systematically organized into sections, each dedicated to a specific focus area of our research. Within each section, we explore and analyze relevant studies, methodologies, and approaches that form the basis of our contributions. By examining prior work, we not only highlight the progress made in the field but also identify unresolved challenges and opportunities for further innovation, thereby positioning our research as a meaningful advancement in both academic and practical contexts.

2.1 Twitter Account Classification

Various studies have focused on classifying Twitter accounts by leveraging different aspects of user behavior, content, and account metadata. A range of features can be extracted for this classification task, with prior works predominantly employing combinations of textual and statistical properties derived from user tweets and account information.

For instance, [1] proposes a method for classifying tweets into two user types—organizations and individuals—using only textual content. By leveraging a rich set of linguistic features and a supervised classifier, their approach achieves high F1-scores of 89% and

87% for English and Spanish tweets, respectively. The study highlights the utility of user type classification in improving event recognition tasks, showing that training separate event models for different user types significantly enhances performance. This work emphasizes the importance of understanding tweet origins for better information filtering and event detection in social media.

Similarly, studies such as [2] and [3] incorporate account metadata alongside user tweets to distinguish between organizational and individual accounts. [2] the distinction between organizational and personal accounts on Twitter, emphasizing its importance for social media studies that focus on human behavior. Using a dataset of over 20,000 accounts, they reveal that organizations, although only 9.4% of accounts, have significantly more extensive social reach and higher follower-to-friend ratios than personal accounts. They introduce a Support Vector Machine (SVM) classifier with an F1-score of 95.5%, leveraging linguistic, stylistic, and structural features, and demonstrate its utility in addressing biases in studies of human-centric Twitter data. The dataset and classifier are publicly available for further research. In their study, [3] present a method for distinguishing individuals from organizations on Twitter using only a single tweet's metadata and profile information. They address limitations in previous work by proposing a character-based convolutional neural network (CNN) and constructing a large dataset using weak supervision, significantly reducing reliance on extensive tweet histories. Their method achieves comparable performance to [2] while requiring fewer data, making it suitable for real-time and large-scale applications. The authors make their dataset, model, and code publicly available, enabling further advancements in Twitter user classification. However, these approaches exhibit certain limitations, including language dependency and the computational burden of data collection and processing noisy, unstructured text data.

Some studies propose extracting statistical features such as the number of favorites, retweets, tweet length, hashtags, and emoticons to address the language dependency of textual analysis. [4] proposes the "POP-MAP" framework for classifying Twitter accounts into three categories: personal, organizational, and personage. Their approach relies on text-based user-generated content and a diverse set of 136 features, including linguistic, stylistic, and platform-specific attributes, such as hashtags, personal words, and links. The study evaluates several classifiers, finding that the

Random Forest model with feature selection achieves the highest performance, with an accuracy of 89.4% and F-measure of 87.9% for English tweets. [5] presents a statistical-based framework for classifying Twitter accounts as individual or organizational, avoiding language-dependent textual content features. Their method leverages metadata of user profiles, post popularity features, and other statistical metrics, achieving high accuracy across different datasets. Using Random Forest as the primary classifier, the approach achieves an F-measure of 95.6% when combined with Synthetic Minority Oversampling Technique (SMOTE) [6]. Although these methods try to overcome the challenge of language dependency, they still require the collection of tweet data and involve complex pre-processing steps.

In scenarios where the primary goal of account classification is to exclude non-personal accounts from a social network graph, relying on tweet content becomes redundant. To mitigate this, other studies utilize the time distribution of tweets or the post frequency of users, bypassing the need for textual processing. [7] introduces a novel approach to classify Twitter accounts as personal, managed, or bot-controlled based solely on the timing of tweets. Using inter-tweet delay and tweet time distributions, they develop Naive Bayes classifiers, achieving 84.6% accuracy for distinguishing personal and managed accounts and 75.8% accuracy for classifying all three types. Their analysis reveals distinct patterns in inter-tweet delays, such as power-law distributions in human-controlled accounts and more structured behavior in bot accounts. [8] proposes a frequency-based method for classifying Twitter accounts into human, bot, and cyborg categories. Their approach relies on statistical features such as average posting intervals, histograms of posts by hour, and other temporal patterns, eliminating the need for textual content analysis. Among various machine learning models evaluated, Random Forest and XGBoost achieved the highest accuracy, with 88% and 84%, respectively, demonstrating the efficacy of their method. However, these approaches share similar disadvantages to prior methods, as they still require access to tweet data.

An alternative direction focuses on the joint representation of profile and network information. For example, [9] addresses the challenge of distinguishing individual accounts from organizational accounts on Twitter by combining text and network features in a novel way. They adapt the Text-Associated DeepWalk (TADW) framework,

incorporating text embeddings derived from sent2vec into network-oriented representations, to create a joint text-network embedding (TNE-s2v). [10] presents a classification model to detect organizational accounts on Twitter using network-based, behavioral, temporal, and spatial features independent of content and language. Their dataset consists of over 7 million tweets from 150,000 accounts in Bangladesh, where less than 1% are organizational accounts such as NGOs, news outlets, and political groups. The sparse logistic regression model achieved the best performance, with a precision of 68.2% and an F1-score of 66.2%. Although effective, graph-based approaches introduce additional computational costs associated with building and analyzing network structures.

To further advance account classification, [11] introduces a multimodal deep neural network, M3, designed for multilingual demographic inference, including classifying Twitter accounts as organizational or individual. Their model processes user profile images, text fields (username, screen name, biography), and employs character-based embeddings fused with language embeddings to handle 32 languages. The model integrates a Long Short-Term Memory (LSTM) network [12] for textual features and DenseNet [13] for processing profile images. Their model is trained on a large, heuristically identified and annotated dataset of 23.86 million samples for the individual vs. organization classification task.

Beyond distinguishing organizational accounts, categorizing users into specific groups offers significant potential for enhancing targeted content delivery. Researchers have explored various methods to classify users based on their interests, allowing personalized recommendations and fostering engagement with content aligned to individual preferences [14, 15, 16]. Others focus on analyzing profile characteristics to cluster users and uncover behavioral patterns, improving the precision of user segmentation [17, 18, 19]. Classification based on political orientation has also been widely studied, providing insights into ideological divides, aiding discourse analysis, and tailoring politically relevant communications [20, 21, 22]. Additionally, identifying user gender has received attention, enhancing demographic analytics and supporting the creation of content strategies sensitive to gender dynamics [23, 24]. These efforts collectively advance the granularity and effectiveness of content delivery by aligning strategies more closely with user-specific traits and preferences.

2.2 Targeted Marketing on Social Media

Personalized advertising has gained significant popularity, driven by the exponential generation of personal data [25, 26, 27]. The term "personalized advertisement" refers to delivering advertisements tailored to a user's interests, typically under the acquisition category. For instance, [28] proposes an advertisement recommendation system for microblogs, addressing the challenges of noisy and sparse data inherent in such platforms. Their method constructs enhanced user profiles by integrating content features from posts (e.g., hashtags, links, captions) with sentiment analysis and user influence derived from follower-followee interactions. A novel aspect of their approach is using Wikipedia Good Articles as background knowledge for matching user profiles with advertisements, enabling a more robust semantic connection. While conceptually related to our work, no direct competitor exists to the proposed pipeline.

[29] describes a framework implemented at Booking.com to optimize the personalization and assembly of user interface (UI) widgets using Multi-Armed Bandit (MAB) algorithms. Their solution allows real-time adaptation to user preferences and environmental changes while maintaining governance over collaborative UI development across teams. The system integrates user feedback directly into the model for continuous learning and adaptivity to optimize landing page user interfaces. Similarly, [30] presents a framework for personalizing landing and home pages at Expedia Group, leveraging reinforcement learning through an MAB framework to improve the traveler experience. Their approach utilizes recent browsing behavior to provide contextually relevant recommendations, supported by machine learning techniques such as Continuous Bag-of-Words (CBoW) for property embeddings and matrix factorization for destination recommendations. These methods significantly enhance the accuracy of recommendations, particularly for repeating visitors, by tailoring suggestions based on user-specific signals like geography and browsing history. The implementation addresses key challenges, including large-scale integration, latency, and real-time personalization, yielding substantial improvements in user engagement and booking efficiency, as demonstrated through A/B testing.

In the automated advertisement-content-generation realm, [31] introduces the Ad-Profile-based Title Generation Network (APTGN), a transformer-based framework

for generating personalized advertisement titles. The model constructs detailed ad profiles using 20 features across basic, stylistic, and current information, incorporating advertiser-specific preferences from historical data. It leverages a feature extraction module with GPT-2 embeddings and integrates these features into a transformer model for title generation. Experiments and online A/B testing demonstrate that APTGN significantly outperforms baseline models. While effective, this approach could benefit from incorporating the contextual message of potential customers to produce genuinely customized titles.

To empirically demonstrate the efficacy of personalized over non-personalized content, [32] conducts a study by analyzing real-world data from Facebook Ads Manager and Google Analytics for a Czech e-commerce platform. Their study uses session-based and user-based segmentation criteria, such as gender, age, and regional behavior, to create tailored landing pages and optimize advertisements for conversion. The results demonstrate that personalized ads outperform non-personalized ads in conversion value, particularly for specific user segments like women aged 25–54, with notable regional differences. The authors highlight the potential risks of over-personalization, emphasizing the need for balanced segmentation to avoid user resistance and maintain effectiveness. However, the segmentation criteria—limited to age and gender—restrict how much the content can be considered genuinely personalized.

Querying individuals based on structured attributes, such as age, gender, and location, is not always feasible, mainly when targeting hard-to-reach demographics. [33] address this challenge by investigating the effectiveness of social media platforms for recruiting hard-to-reach populations, focusing on users of direct-to-consumer genetic testing (DTC-GT) services who utilize third-party interpretation tools. The study compares recruitment outcomes across Facebook, Twitter, and Reddit using paid campaigns with tracking metrics (click-based vs. conversion-based) and unpaid community posts. Results reveal that conversion-based tracking on Twitter was the most cost-effective approach, while Reddit provided a younger and more male-dominated sample. Their findings highlight the potential of tailored social media strategies for survey recruitment, emphasizing demographic variations across platforms and tracking methods. Despite their contributions, limitations inherent to web-based survey recruitment hinder the acquisition of detailed user information.

In the context of social engagement and interaction, [34] provides a comprehensive analysis of the factors driving four types of social engagements on Twitter: likes, retweets, replies, and quotes. They employ diverse features, including user behavior, tweet attributes, and text semantics derived from BERT embeddings, to predict engagement using over 14 million multilingual tweets. The study reveals that user engagement is primarily driven by text semantics, irrespective of the tweet author. Furthermore, combining semantic analysis with similarity-based user features improves the predictability of social engagements. Notably, the study challenges the reliance on collaborative filtering, demonstrating that users do not always follow the behaviors of previously engaged users. This finding has critical implications for marketing, where social engagement is a proxy for advertisement interaction. The authors emphasize that extracting context from a potential customer's message yields more effective results than relying solely on historical user preferences.

In natural language generation (NLG), sentence arrangement is crucial in producing coherent, readable text. Improper ordering can significantly degrade text quality, an issue widely studied in multi-document summarization [35, 36]. Scholars have explored methods to create summary paragraphs considering topical relatedness, chronology, and sequence. Additionally, researchers have tackled the problem of restoring paragraph coherence from unordered sentence sets [37, 38]. In contexts where generic sentences are generated, thematic proximity between sentences must be prioritized to ensure paragraph coherence, particularly in applications such as landing page generation.

Long-text generation further highlights the importance of structuring content. Human writers typically plan content organization before refining each section's surface form. To mimic this process, [39] decomposes text generation into a series of dependent sentence-structure tasks, addressing local coherency through context-aware generation. A globally coherent paragraph is then constructed via a greedy approach. They propose the Planning-based Hierarchical Variational Model (PHVM) to address challenges in generating long and diverse texts from structured data. The model incorporates a two-stage hierarchical approach, first performing high-level planning by grouping input attributes into clusters and then generating sentences conditioned on these clusters. A key innovation is the integration of global and local latent variables

to capture diversity and coherence, enabling contextually appropriate text generation. In contrast, [40] proposes PermGen, a framework for generating paragraphs with improved diversity by permuting sentence orders during generation rather than adhering to a fixed sequence. The model incorporates hierarchical positional embeddings to handle sentence permutations and utilizes π-SGD optimization to ensure convergence across all possible sentence orders. Experiments across three tasks—story generation, news generation, and paper abstract generation—demonstrate that PermGen outperforms baseline models like BART, GPT-2, and T5 in both diversity and accuracy. The framework shows particular strength in producing diverse, semantically rich paragraphs, addressing the limitations of traditional left-to-right generation models. Both approaches demonstrate a strong correlation between paragraph quality and sentence arrangement.

Topic modeling remains a widely adopted technique for identifying latent themes in large text corpora. While probabilistic topic models are frequently used [41], pre-trained word embedding models offer competitive performance with reduced computational overhead. For example, [42] proposes an alternative to probabilistic topic models, such as Latent Dirichlet Allocation (LDA), by clustering pre-trained word embeddings for topic extraction. Their approach incorporates weighted clustering and reranking techniques to enhance topic coherence and relevance. By benchmarking various combinations of word embeddings and clustering algorithms, they demonstrate comparable performance to LDA while achieving lower runtime and computational complexity. Additionally, they show that dimensionality reduction through PCA can significantly speed up the process without sacrificing topic quality, making their method a practical choice for large-scale document analysis. They report that BERT (Bidirectional Encoder Representations from Transformers) [43] achieves the best average performance, followed closely by GloVe (Global Vectors for Word Representation) [44].

For targeted topic modeling, [45] introduces SeededLDA, an extension of LDA that incorporates lexical priors to guide topic models toward discovering user-relevant topics. The model allows users to provide seed words representing specific topics, which are then used to bias both topic-word and document-topic distributions. This approach addresses the limitations of traditional LDA, which often fails to identify rare

or underrepresented topics, by integrating user knowledge to improve the semantic coherence and relevance of topics. Extrinsic evaluations on document clustering tasks demonstrate significant performance improvements, highlighting the utility of incorporating lexical guidance into unsupervised topic modeling. Recent advancements such as BERTopic have further improved unsupervised topic modeling, demonstrating competitive results on large-scale datasets with enhanced performance and scalability [46].

2.3 Aspect-Based Sentiment Analysis

Substantial progress has been made in recent years in Aspect-Based Sentiment Analysis (ABSA), driven primarily by deep learning approaches and the use of large-scale pre-trained language models. Early works leveraged transformer-based architectures such as BERT, yielding significant improvements in performance across ABSA subtasks.

[47] explores the potential of contextualized embeddings from pre-trained language models, particularly BERT, for End-to-End Aspect-Based Sentiment Analysis (E2E-ABSA). Their approach reformulates E2E-ABSA as a sequence labeling task and investigates multiple downstream neural architectures, including linear classifiers, GRUs, self-attention networks, and CRFs. Experimental results on benchmark datasets demonstrate that even simple models fine-tuned with BERT representations achieve notable results, underscoring the effectiveness of context-aware embeddings. The study also highlights BERT's robustness against overfitting, establishing a reliable benchmark for E2E-ABSA tasks. [48] propose a dual-MRC framework that effectively handles all subtasks in aspect-based sentiment analysis (ABSA), including aspect term extraction (ATE), opinion term extraction (OTE), and sentiment classification. This approach leverages two BERT-based models for joint training, treating ABSA as a machine reading comprehension task. These outputs are then combined to produce the final sentiment predictions. The framework's efficiency stems from its span-based extraction and classification, making it suitable for comprehensive ABSA tasks in a unified model. Similarly, [49] introduces a span-level approach, Span-ASTE, to enhance Aspect Sentiment Triplet Extraction (ASTE). The

method explicitly captures interactions between entire spans of aspect targets and opinion terms, overcoming the limitations of previous word-level methods. By integrating a dual-channel span pruning strategy supervised by ATE and OTE, the framework achieves computational efficiency and improved pairing of target and opinion spans. Span-ASTE significantly outperforms previous approaches in extracting triplets across benchmark datasets, especially those with multi-word targets or opinions.

On the other hand, sequence-to-sequence architectures have also been explored for ABSA. [50] introduces an approach to address all seven subtasks of ABSA using a unified generative framework. By redefining each subtask as a sequence generation problem, the authors leverage the sequence-to-sequence pre-trained model BART to provide an end-to-end solution. This approach significantly advances the unification and practical implementation of ABSA subtasks. [51] build upon this work by introducing improvements in data preprocessing and architectural modifications, resulting in further performance enhancements. By modifying BARTABSA's outer layers and leveraging its generative capabilities, the authors demonstrate significant improvements in both the Laptop-ACOS and Restaurant-ACOS datasets compared to prior benchmarks. The study highlights the model's flexibility and ability to unify sentiment analysis subtasks without altering its core algorithm.

More recently, InstructABSA [52] diverges from the traditional BERT-based approaches by incorporating instruction-tuning approach for various ABSA subtasks. By fine-tuning the Tk-Instruct model with task-specific instructions, including examples with positive, negative, and neutral sentiments, it outperforms state-of-the-art models across several datasets, such as SemEval 2014-2016. Notably, InstructABSA substantially improves tasks like ATE and aspect sentiment pair extraction (ASPE) while maintaining strong sample efficiency, requiring only 50%

In addition to transformer-based methods, alternative techniques have been proposed to capture more intricate relationships between aspects and sentiments. [53] proposes the Disentangled Linguistic Graph Model (DLGM) model to improve performance and explainability in ABSA. It introduces a disentangled linguistic representation learning module to capture finer linguistic features and a cross-linguistic routing

mechanism to enhance transparency. The model leverages graph neural networks (GNNs) with specific linguistic properties, such as part-of-speech tags and syntactic dependencies, to provide detailed insights into sentiment predictions. Similarly, [54] presents an approach to ABSA by leveraging disentangled attention mechanisms within the DeBERTa model. This method explicitly separates semantic and syntactic features, allowing for enhanced representation of textual data in ABSA tasks.

2.4 AI-Based Content Generation

The application of AI in content generation, manipulation, and advertising has drawn significant attention in recent years, revolutionizing industries such as social media and marketing. By leveraging AI's capability to analyze extensive datasets, identify patterns, and produce contextually relevant content, these advancements have reshaped how content is created, consumed, and personalized.

AI algorithms have transformed the way content is created and consumed on social media platforms. Research highlights the effectiveness of generative AI in automating creative tasks, including social media creatives and advertising campaigns. [55] explores the effectiveness of AI in creative tasks, particularly in generating social media visuals. Through extensive human evaluation, the study finds that AI-generated creatives often outperform human-designed ones in engagement metrics like likes and comments, particularly for short text descriptions and content involving animals or products. Similarly, [56] highlights the increasing threat posed by AI-generated disinformation in user-generated content, particularly in marketing research. It demonstrates AI's ability to fabricate realistic disinformative content, quantifies its disruptive effects on marketing analytics frameworks, and evaluates advanced detection models, finding that standard methods often fall short. The study also advocates for a multi-faceted approach combining algorithmic solutions, human oversight, and regulatory updates to counteract this growing issue, offering a foundation for future research and policy-making in this domain.

Personalization is another domain where AI excels. The study by [57] investigates the prevalence and usage of AI-generated profile pictures on Twitter, focusing on nearly

15 million profile images. It identifies that 0.052% of these images are artificially generated and highlights their usage in spam, disinformation campaigns, and political amplification. The authors employ a multi-stage detection pipeline and analyze patterns of coordinated inauthentic behavior among the associated accounts, revealing their involvement in contentious topics like COVID-19 and political discourse. [58] explores the ethical challenges of AI-driven digital manipulation, emphasizing its pervasive and opaque nature. It conceptualizes "digital manipulation" as influence mediated through AI and digital technologies, intentionally bypassing reason to create an asymmetry of outcomes between manipulators and users. The work highlights the risks to cognitive liberty, autonomy, and social polarization posed by practices like microtargeted advertising, social media algorithms, and deepfakes. The author argues that these technologies necessitate robust ethical frameworks and regulatory measures to mitigate manipulation while safeguarding autonomy and mental self-determination.

Various researchers have critically examined the manipulation potential of AI in social media. [59] investigates how prompt engineering can manipulate large language models into generating emotionally charged disinformation, highlighting the risks of AI misuse in spreading misinformation. The authors demonstrate that polite emotional prompts significantly increase the likelihood of these models producing disinformation, while impolite prompts reduce compliance. It uses a corpus of 19,800 generated posts across various public health and societal topics. The research highlights the susceptibility of LLMs to emotional exploitation and underscores the need for ethical AI design. [60] further investigates the challenges in defining and mitigating AI-driven manipulation, proposing ethical safeguards to prevent advertising and social media exploitation. The authors explore the critical topic of manipulation in artificial intelligence systems, particularly in domains like recommender systems and language models. The study introduces a framework for understanding manipulation through four axes: incentives, intent, covertness, and harm. The authors highlight challenges in defining and operationalizing these concepts, emphasizing that manipulation can emerge without system designers' explicit intent. The work underscores AI manipulation's risks to human autonomy and advocates for precautionary measures, such as transparency and auditing, to mitigate potential harms.

Integrating generative AI in advertising has also raised questions about ethical boundaries. [61] explores how artificial intelligence enhances contextual advertising by optimizing ad placements through a deeper semantic understanding of media content. It highlights the shift from reliance on personal data, driven by privacy regulations like GDPR, towards contextual targeting based on content relevance. The study discusses the role of AI in identifying key context factors, such as topical applicability, affective tone, and consumer engagement, to improve ad effectiveness. However, it also raises concerns about ethical risks, including discriminatory targeting and manipulative advertising, emphasizing the need for transparency and self-regulation in AI-driven systems.

AI's influence extends beyond creative tasks and manipulation. [62] analyzes the impact of AI-generated content (AIGC) on social media dynamics using a dataset of 15.2 million posts, including 2.4 million AI-generated images from Pixiv. The study reveals a 50% increase in artwork submissions post-AIGC adoption but notes an absence of proportional growth in user engagement metrics such as comments or views. The findings also illustrate shifts in content diversity and consumption patterns, with mid-tier AI-generated content outperforming human-created equivalents in popularity.

Despite its benefits, the misuse of AI for disinformation and manipulation remains a pressing concern. [63] argues that AI's capacity for generating targeted political misinformation poses significant challenges to democratic processes, necessitating robust countermeasures. The author discusses stricter standards against AI deception than human norms, proposing that AI outputs leading users away from "semi-ideal" beliefs or choices be classified as deceptive or manipulative. The study introduces measures like "extreme transparency" and defensive systems to counter these risks by labeling AI-generated content with contextual information and ensuring accountability. The paper also explores the potential long-term dangers of agentic AI systems capable of autonomous deceptive behavior. Similarly, [64] highlights the growing prevalence of fake news, which exploits traditional news media's credibility while undermining public trust. It defines fake news as fabricated information designed to mimic news media content without adhering to editorial standards or organizational processes for accuracy. The study emphasizes the role of social bots in amplifying

fake news dissemination and calls for a multidisciplinary approach, including structural and educational interventions, to combat its societal impacts. The primer [65] on artificial intelligence, deepfakes, and disinformation explores how advancements in AI have enabled the creation of hyper-realistic synthetic content like deepfake videos, voice cloning, and generative text. These technologies amplify the potential for malicious disinformation campaigns, targeting elections, social cohesion, and trust in institutions. The report also highlights ongoing efforts to counteract these threats, including improved detection technologies, regulatory frameworks, and media literacy programs, while emphasizing the need for coordinated policy and technological solutions to address emerging risks effectively.

LLMs' pervasive challenge of mitigating toxic content generation garners significant attention in recent research. [66] proposes a dual-model strategy, training one GPT-2 model specifically to produce toxic content and another for conventional use. By re-ranking next-word predictions against the toxic model's outputs, they successfully reduce harmful generations. Building on this, [67] introduces a paired dataset containing toxic and non-toxic text versions, fine-tuning LLMs on this curated corpus. Their approach, validated through human evaluations, achieves a notable reduction in toxicity, emphasizing the critical importance of data curation in addressing this issue. [68] further advances the field by exploring text generation control in response to toxic prompts. Their method employs an auto-regressive model that quantizes and adds noise to data, enabling the model to learn to reconstruct the original input. Additionally, reinforcement learning with token-level feedback is used to suppress toxic outputs further. Collectively, these studies highlight a dual approach to combating toxic content: carefully curating training datasets to define undesirable outputs and robust post-generation filtering mechanisms to eliminate residual toxicity.

Based on the aforementioned issues, it is evident that AI's rapid advancement and integration in content generation bring both remarkable opportunities and significant challenges. While AI has transformed industries by automating creative tasks and enhancing personalization, its susceptibility to misuse for disinformation, manipulation, and the generation of toxic content underscores the urgent need for ethical and technical safeguards. There is a compelling demand for programmable AI systems that exhibit advanced capabilities and prioritize humanizing principles. Such systems

must be designed to actively mitigate biases, toxicity, and manipulative tendencies while fostering transparency, accountability, and alignment with ethical standards. Moving forward, the development of AI must incorporate these human-centric values to ensure its application remains a force for societal benefit rather than harm.

CHAPTER 3

TWITTER ACCOUNT CLASSIFICATION USING ACCOUNT METADATA: ORGANIZATION VS. INDIVIDUAL

Social media has become an essential part of life, not only for humans but also for organizations and establishments. Nowadays, the world is in an era of connectivity and accessibility. Fast and easy access to knowledge, sharing information, and connecting the world became fundamental needs for society. Several social media platforms emerged to meet these demands, and one of the biggest platforms is Twitter. Twitter is a microblogging system that allows users to share relatively short content. Twitter also contributes a wide variety of research opportunities in various areas with the data produced in it [69].

Twitter provides rich content for various aspects of social media analysis studies, such as text mining, graph structure, or statistical analysis. However, the data needs preprocessing and cleaning since personal accounts are not structured, and there are also many fake, bot, and corporate accounts. Manual annotation of datasets is costly and biased since annotators' decisions might be subjective. Therefore, various research attempts try to make account type detection automatic [70]. This automated account classification is a necessary prior step to improve the quality of various social media analysis tasks, such as building social media graphs or analyzing message texts. Space and time complexity are one of the biggest challenges in building social network graphs. Thus, account elimination is significant in reducing this complexity [71, 72, 73].

One of the popular research subjects that also requires tweets is the sentiment analysis [74, 75, 76, 77], which is widely used for identifying user's reviews of entities to capture the trends. Similarly, event-related Twitter posts are trendy research top-

ics [78, 79, 80, 81]. Tweets themselves are the primary source of information for such research. In similar studies, individual accounts are needed, and organizational accounts must be excluded.

Our study focuses on the same problem and aims to automatically determine organizational and individual accounts. We propose a powerful approach to classify Twitter accounts as individual vs. organization, using limited metadata of the accounts. Organizations, institutions, news, or entertainment accounts are labeled as *organization*. On the other hand, *personal* accounts are the accounts that reflect a human's emotion, thought, or behavior, so fake user accounts can also be personal accounts.

Table 3.1: Humanizr and Demographer dataset statistics.

Dataset Name	Original Accounts	Collected Accounts	Collection Percent
Humanizr	18,922	17,790	94%
Demographer	227,277	214,236	94%

As one of the previous studies on the same problem, in [2], the last 200 tweets posted from a given account are analyzed for account classification. Following a similar approach, the work in [3] adopts a single tweet to feed into a character-based convolutional neural network. Unlike these two, we concentrate on only the account metadata; thus, in our approach, the data collection, preprocessing, and classification stages require much less effort than the related methods. Furthermore, the metadata of protected accounts is also publicly available; thus, it is possible to classify the protected accounts. The proposed method is publicly available as a software tool.

Today, many Twitter accounts contain texts written in multiple languages and dialects. Therefore, multi-lingual support is needed for the account classifiers that analyze message content. Our proposed model does not depend on a single language. It encodes the textual properties using Byte-Pair Encoding [82].

Due to the nature of Twitter use, the number of individual and organization accounts is inherently unbalanced, such that the individual accounts constitute the majority [11]. To overcome this deficiency for model construction, we have followed the procedure used in the studies given in [3] and [2], where the dataset is arranged as balanced,

unbalanced, and full.

3.1 Method

This section outlines the steps taken to prepare and process the data for the classification task. We aimed to create a streamlined and replicable methodology by leveraging publicly available datasets and open-source tools. The data collection process, described in detail below, ensures a comprehensive approach to capturing relevant features while minimizing unnecessary complexity in the preprocessing phase.

3.1.1 Data Collection

Our method's data collection step is less laborsome than those in similar studies. Previous studies given in [3] and [2] provide publicly available data sets, Humanizr and Demographer. The repository[1] given in [3] provides the Demographer dataset as a text file containing Twitter account IDs and labels for each account. The authors also provide the Humanizr dataset, which they collected in 2018, in the same format. Using Twitter account ID, we have fetched the profile metadata of Twitter accounts using the open source Twitter scraping tool 'Twint' [83]. Since we use only the profile information, the dataset also includes protected accounts. We fetched 17,790 Twitter account profiles from the Humanizr data set and 214,236 accounts from the Demographer. Table 3.1 provides the number of successfully fetched accounts in the datasets. Original numbers are taken from the study in [3]. The original and collected counts differ from closed or suspended Twitter accounts. Table 3.2 shows the profile metadata for a sample account comprising 16 attributes. Among these attributes, four textual and five numeric attributes are used for model construction. We excluded attributes with low potential to contribute to the classification, such as ID and join date. The avatar attribute is also taken out as it involves image processing, which may increase the processing complexity. No additional feature filtering method was applied to the data.

[1] Knowles et al. (2016). Demographer Dataset [online]. Website https://bitbucket.org/mdredze/demographer/src/master/data/ [accessed December 2024].

Table 3.2: Twint provides 18 different attributes for user profiles. The attributes used in this study are marked in bold.

Profile Attributes	Description	Type
id	Unique identifier of user	integer
name	Name of the user	text
username	Unique screen name	text
bio	Description of the account	text
location	User-defined location	text
url	A url provided by user	text
join date	UTC date of account creation	text
join time	Time of account creation	text
tweets	Number of tweet of the account	integer
following	Number of users this account following	integer
followers	Number of users this account has	integer
likes	Number of tweets user has liked	integer
media	Number of media in the tweets	integer
private	True when account is protected	boolean
verified	True when user has verified account	boolean
avatar	URL pointing to the user's profile image	text

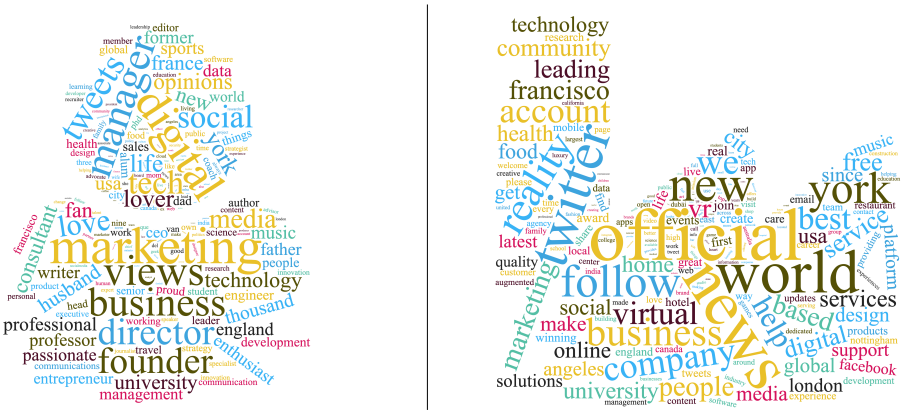


Figure 3.1: Word cloud of Demographer dataset, containing top 250 occurred words in individual (left) and organization (right) accounts.

3.1.2 Data Analysis

The collected Humanizr dataset consists of 17,790 user accounts, 16,012 of which are labeled as individuals, and 1,778 of them are labeled as organizations. In the Demographer dataset, there are 214,236 accounts, 185,224 labeled as individuals, and 29,012 labeled as organization. Since the Demographer dataset comprises the Humanizr dataset, the figures are based on the Demographer dataset in the following sections. The source codes, the dataset, and the interactive versions of the given figures used in this study are publicly available[2].

We selected *name*, *bio*, *location*, and *url* attributes as text features and *tweets*, *following*, *followers*, *media*, and *likes* attributes as numeric features for the classification. Figure 3.1 shows a high overlap between used words for each account type.

Figure 3.2a visualizes the relationship between the number of followers and followee per account. The individual accounts have a high number of followees, but few followers try to increase their visibility by using the follow-back strategy. The majority of this group is entrepreneurs. In contrast, accounts with a high number of followers and fewer followees are usually owned by the people who are recognized by the crowd with their success in their expertise areas. Similarly, organizational accounts of large companies have a high number of followers and few followees. Examples include newspapers, news agencies, social media companies, and the entertainment industry's well-known brands. Non-profit or charity foundations are not concerned about following-back accounts since they aim to increase their visibility and reach out to the people.

The individual accounts in Figure 3.2b with more than 500K tweets do not have many followers. Still, they continue to tweet consistently. Similar to Figure 3.2a, high-followers less-tweets points are celebrity accounts. They do not aim to be active with posting tweets since their reputation in the real world is high enough. The organization accounts having high tweet counts are the customer support accounts of companies. Customers tweet by mentioning their account names while expressing

[2]Çetinkaya et al. (2021). Twitter Account Classification Using Account Metadata: Organization vs. Individual, *Source and Dataset* [online]. Website https://github.com/tweetpie/twitter-account-classification [accessed December 2024].

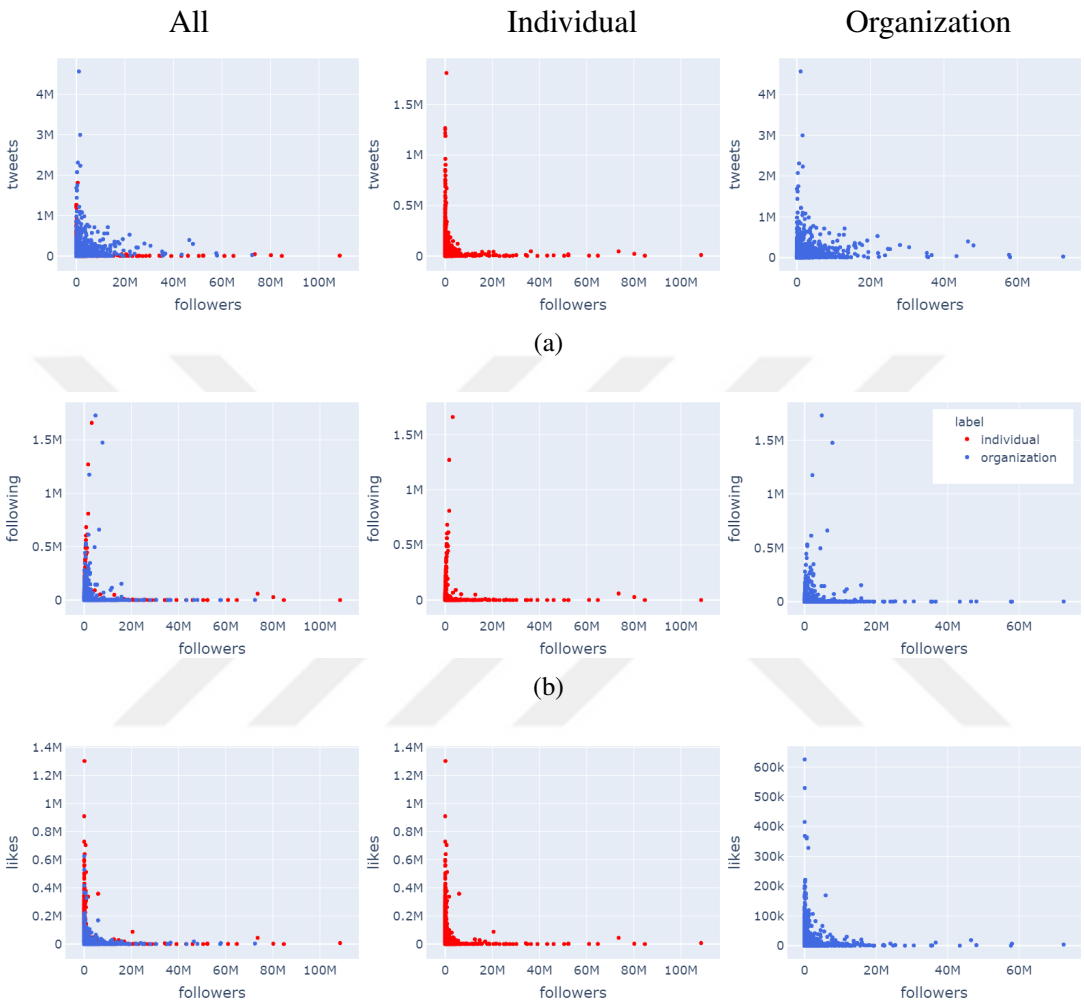


Figure 3.2: The distribution of individual and organization accounts in the Demographer dataset comparing the follower count vs. (a) followee, (b) tweet, and (c) like counts. Note that the first chart in each row is for the whole collection of individual and organizational accounts.

problems with the product or service. By nature, these accounts do not have that many followers. The organization accounts with so many followers tend to tweet less. The pioneer enterprises adopt this behavior mostly. Organizations tweet more than individuals independently of their follower counts. Since professionals mostly manage those accounts, they have a consistent tweet-sharing behavior compared to the individuals.

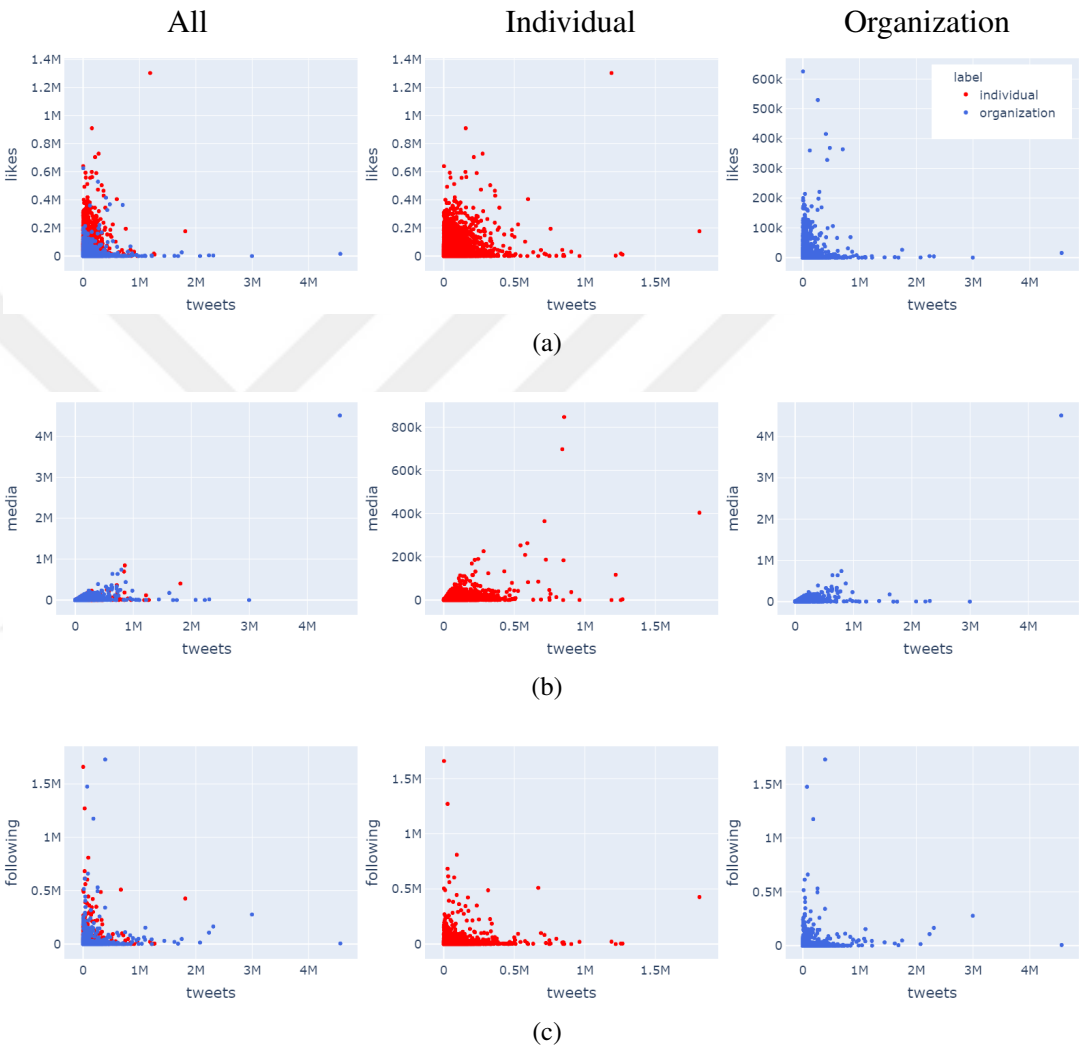


Figure 3.3: The distribution of individual and organization accounts in the Demographer dataset comparing tweet count vs. (a) like, (b) media, and (c) following counts. Note that the first chart in each row is for the whole collection of individual and organization accounts.

Figure 3.2c shows that individuals with fewer followers use Twitter as regular users, consistently following other accounts and trend topics. This situation makes their

likes count higher than the others. Individuals with high followers tend not to like others' tweets. The reason might be that they feel uncomfortable since their likes are exposed to the mass. The main aim of organizations with few followers is to increase their visibility by liking other tweets. The accounts following this strategy are mostly small or medium-sized enterprises and organizations trying to build a community. In contrast to 3.2a, charities and non-profit organizations take up less space in the group of organizations with a high number of likes. Twitter accounts of big companies have a similar attitude to being neutral, considering the company's interests. Organizations like tweets as half as individuals do on average.

As shown in Figure 3.3a, individuals have more likes than organizations in general. Celebrity accounts are close to the origin, where they usually tweet and like less. At some point, the count becomes acutely fewer while the tweet count increases. Support accounts have more tweets but fewer likes, as expected. The balance between tweets and likes is more consistent for individuals than organizations. The organizations with a few tweets but many likes usually do not post original tweets but retweet or like other accounts' content.

The diversity of media use in individuals' tweets is higher than in organizations. The organization accounts with media in almost all of their tweets, mostly propaganda or ad accounts. The organization accounts with a high number of tweets and fewer media in Figure 3.3b are the support accounts. They usually reply to the clients asking for help with plain text.

Individuals following so many accounts do not frequently tweet, as displayed in Figure 3.3c. They aim to gain followers with the follow-back strategy. In contrast, the ones who tweet a lot do not follow that many users. These two groups pursue different strategies to become famous. Nonetheless, accounts with less than 500K tweets have a more consistent number of followees. Similar behavior can be seen in the organization's accounts as well.

Figures 3.2 and 3.3 demonstrate that numeric features for organization and individual accounts do not give explicit clues to distinguish them. The accounts having similar goals are gathered together, as explained in this section. However, the numeric features do not have specific organization vs. individual classification information.

Therefore, we used their textual features accompanied by numeric ones to classify.

Table 3.3: The statistics for the word lengths in textual features in Demographer dataset.

	Average	Standard Deviation	Max
Bio	11	5.97	105
Name	1.97	0.74	33
Location	1.55	1.96	34
Url	2.48	1.53	48

Table 3.4: The statistics of the numerical features in Demographer dataset.

	Average	Standard Deviation	Max
Tweets	22,982.40	67,810.64	2,706,830
Following	3,007.94	32,085.57	1,196,529
Followers	345,570.94	2,718,346.26	108,539,399
Likes	8,790.79	28,226.66	617,782
Media	5,256.03	25,707.56	752,000

Table 3.3 and Table 3.4 show the statistical summary of the features. For the textual features, average, standard deviation, and max word counts are investigated to apply correct padding in the preprocessing step. The same statistical investigation is applied to numerical features. However, outliers with extremely high values for the features affect the overall average, standard deviation, and max values, as shown in Table 3.4. Therefore, these statistics do not give valuable insights into the numerical features in the preprocessing phase.

The Demographer dataset includes Humanizr dataset accounts; thus, the data analysis section focuses on the Demographer. We considered that the URL, location, and bio fields might differ according to account type, and hence they can provide a clue for the account classification task. Figure 3.4 shows the number of accounts that do not have these attributes for both classes. The missing URL attribute percentage among organization accounts is 7%, but among individual accounts is 21%. These statistics might indicate that if a URL attribute is not specified for an account, it is more likely an individual account than an organization.

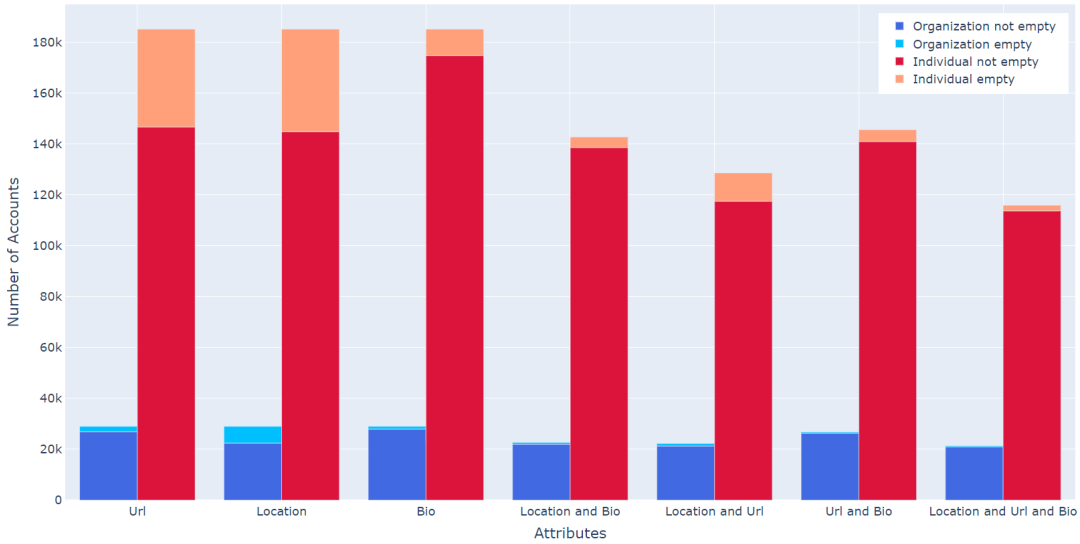


Figure 3.4: Analysis on accounts for existence of URL, location and bio attributes in Demographer dataset

3.1.3 Data Preprocessing

In the preprocessing step, we applied basic data-cleaning techniques. The numerical values in the textual data are converted into textual equivalents to facilitate the model training. We removed all punctuation and made the whole text lowercase. Stopwords for the languages used in the dataset are removed from the account texts. Normalization has been applied for the numeric attributes, and the values are scaled into the range [0-1].

There are accounts in various languages, making finding a standard embedding technique to prepare texts as input to the neural network challenging. Additionally, since Twitter users do not focus on grammar rules or language structure in their account information, some words are concatenated in these metadata texts. Therefore, pretrained models for word embeddings, such as GloVe [44] or word2vec [84], could not work directly for our dataset because they do not include these concatenated words in their corpora. We used Sentencepiece [85] to extract the tokens from compound words.

We have also observed that some sentences include words from different languages together. The Sentencepiece tool is a multilingual unsupervised tokenizer and deto-

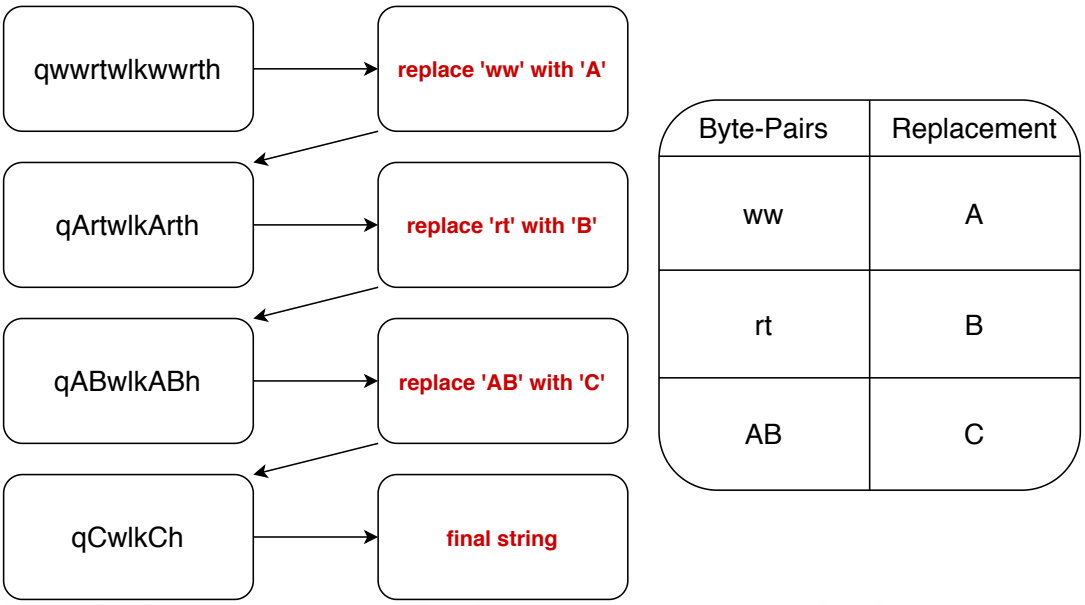


Figure 3.5: Original Byte-Pair Compression Algorithm

kenizer. It supports the Byte-Pair Encoding segmentation algorithm to overcome unknown word problems [82]. Unknown words can be represented with unique tokens like <unk> in popular word embeddings like word2vec or GloVe, but it is not a viable solution. Different approaches to solving this problem exist, such as subword segmentation, which reconstructs a word's meaning from its parts. Byte-Pair Encoding is an unsupervised subword segmentation method that starts with a sequence of symbols and iteratively merges the most frequent symbols to generate a new symbol. The origin of Byte-Pair Encoding comes from the [86], and it is presented as a data compression algorithm that works by replacing it with a byte that does not appear in that data as can be seen in Figure 3.5.

The study in [82] adopts this algorithm and modifies it slightly to provide a solution to the subword tokenization problem. The original implementation's main difference is combining the most recurring characters or words instead of replacing them with another byte for compression. For example, in English corpora, the most recurring characters 'h' and 'e' are combined into 'he,' 't' and 'h' are combined into 'th', then 't' and 'he' are combined into 'the' to generate a new word.

We have used MultiBPEmb presented in [87], a collection of multilingual subword segmentation models and pre-trained subword embeddings. It can tokenize concate-

Table 3.5: Sentencepiece tokenization results for some examples.

Language	Text	Tokens
en	defenceandtechnology	['_defence', 'and', 'technology']
	ventureforamerica	['_venture', 'for', 'america']
	lifeinpleasantville	['_life', 'in', 'pleasant', 'ville']
	peteforamerica	['_pete', 'for', 'america']
de	diewahrheit	['_die', 'wahr', 'heit']
iw	רעצדוע	['_דוע', 'רעצ']
ru	степеньпромышленногопроизводства	['_степень', 'промышлен', 'ного', 'произ', 'водства']
es	bailarconmusica	['_bailar', 'con', 'musica']
fr	chargédecommunication	['_chargé', 'de', 'communication']
it	nontihoamore	['_non', 'ti', 'ho', 'amore']

nated words in 275 languages and provide 300-dimensional subword embeddings for different Byte-Pair Encoding vocabulary sizes. MultiBPEmb uses SentencePiece to learn Byte-Pair Encoding subword segmentation models and GloVe to train subword embeddings. Table 3.5 shows the tokenization results of several examples. Each token has a 300-dimensional embedding vector representation.

3.1.4 Classification

Our proposed model consists of two parallel models: multi-layer perceptron (MLP) for numerical features and bi-directional long short-term memory (LSTM) for textual features. Figure 3.6 details the model's architecture.

MLP consists of multiple perceptrons echeloned into multiple layers. A single perceptron takes numeric inputs with the corresponding weights to emit these values' summation to the next layer's neurons. The signal emitted from the output layer is used to predict the class of the given sample. The architecture may vary with the number of layers, neurons, and activation functions applied to the neurons.

LSTM is a type of recurrent neural network that takes the input as a temporal sequence [12]. The neurons inside the LSTM are called memory cells, where the state is stored. Each unit in the sequence is fed one by one, and the state of the cell is updated throughout the sequence. The cell encapsulates the representation encoded by the network when the sequence is fully consumed. Like traditional neurons, they may have activation functions at the end of each node to map the input with non-linearity.

We have applied an exhaustive search for hyperparameters while designing the model

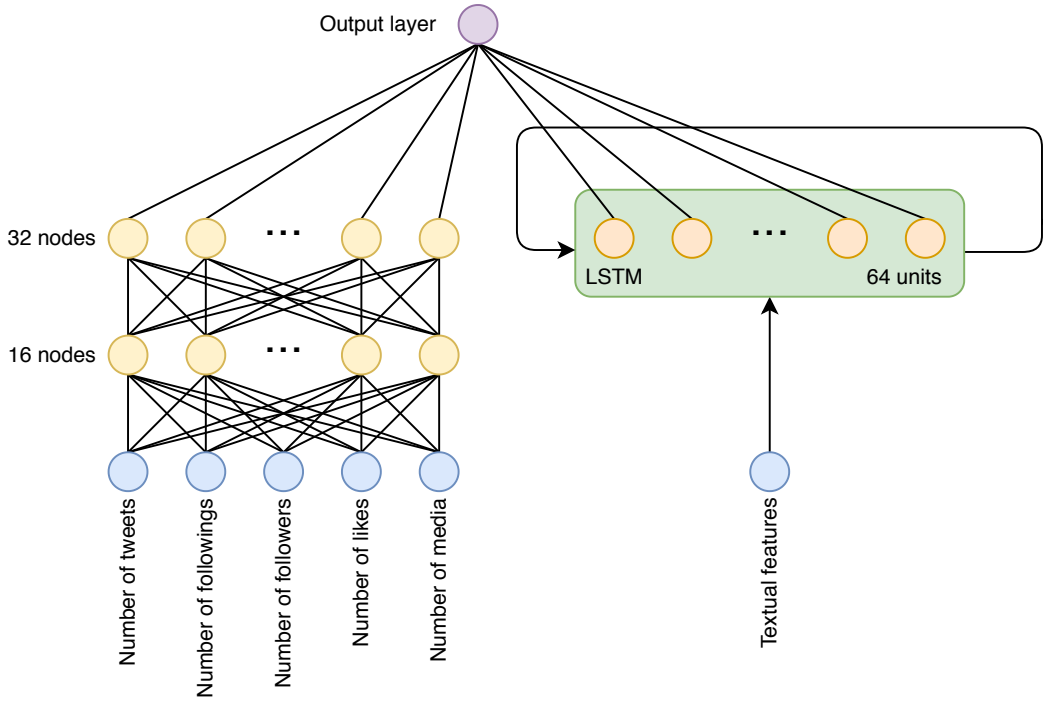


Figure 3.6: Multi-layer perceptron (MLP) for numeric features and long short-term memory (LSTM) for textual features are used.

by splitting the training set into validation and training without using the test set. One to four layers of MLP are tested with 16, 32, and 64 nodes inside the hidden layers. Similarly, 32, 64, and 128 sequence lengths are tested for the LSTM part. The configuration that provides the highest accuracy scores in validation analysis is applied in the final model.

We feed the MLP model with five numeric features: the number of tweets, followings, followers, likes, and media. It has two hidden layers with 16 and 32 hidden nodes correspondingly. Both layers have a rectified linear unit (ReLU) activation function at the end. Bi-directional LSTM has 64 units inside, and 300-dimensional embedding vectors are fed. We set the maximum sequence length to 64. A dropout layer is connected to LSTM for regularization. Random nodes are removed at each iteration to prevent the model from over-fitting. The output of the last MLP and LSTM layers are concatenated before feeding it to the output layer. The activation function of the output layer is the sigmoid function for binary classification. The loss function for our model is binary cross-entropy. The model uses the Adam optimizer to adjust weights. We used a decaying learning rate, starting from 10^{-3}. We selected a batch size of 128

for the training process.

3.2 Experiments and Evaluation

We have implemented our model using the Tensorflow library in Python, which is successfully applied in similar tasks [88]. The experiments are conducted under k-fold cross-validation where $k = 7$. We adopted four evaluation metrics for classification performance with a weighted average: accuracy, precision, recall, and F-measure. Three studies focus on the same classification problem with different approaches [2, 3, 11]. Their evaluation also differs; while the studies in [2] and [11] provide recall and accuracy, [3] reports only overall accuracy. Since we have four different evaluation criteria encapsulating previous techniques, we take these studies as baselines and compare them with our results.

Table 3.6 shows the previous studies' evaluation results compared to our proposed model. We call the model names the same as the data on which they are trained. The table compares baseline models with the proposed model in a paired manner. The results of Humanizr [2], Demographer [3], and M3 [11] models are given as reported in the original papers. However, each model is applied to different versions of the datasets, such that the Humanizr model is applied to the balanced, unbalanced, and full version of the Humanizr dataset. On the other hand, the Demographer model is applied to the balanced and full version of the Demographer dataset. The balanced version of the data set includes an equal number of individual and organization accounts. The number of accounts is determined by undersampling the majority class. They train and test on the balanced dataset, leading to a different model and evaluation results. Balancing the dataset decreases the total account number (2 x the number of organizations). To make a fair comparison, in the Humanizr evaluation, the authors also prepare an unbalanced dataset by keeping the organization and individual account ratio the same as the original dataset but the overall account number the same as the number of accounts in the balanced dataset. Hence, we have followed the same steps to generate the same experiment environments by preparing balanced and unbalanced versions of the Humanizr and balanced versions of the Demographer dataset.

Table 3.6: Performance measurements of each model under k-fold cross-validation.

Model	Dataset	Class	Recall	Accuracy
[2]	Humanizr (Full)	Organization	66.0%	**95.5%**
		Individual	**98.6%**	
Proposed Model	Humanizr (Full)	Organization	**71.8%**	95.0%
		Individual	97.6%	
[2]	Humanizr (Unbalanced)	Organization	**58.6%**	**94.1%**
		Individual	**98.2%**	
Proposed Model	Humanizr (Unbalanced)	Organization	57.6%	92.5%
		Individual	96.9%	
[2]	Humanizr (Balanced)	Organization	**89.4%**	**88.8%**
		Individual	88.2%	
Proposed Model	Humanizr (Balanced)	Organization	85.6%	87.6%
		Individual	**89.6%**	
[3]	Demographer (Full)	Organization	-	94.6%
		Individual	-	
Proposed Model	Demographer (Full)	Organization	**90.4%**	**97.4%**
		Individual	**98.5%**	
[3]	Demographer (Balanced)	Organization	-	85.8%
		Individual	-	
Proposed Model	Demographer (Balanced)	Organization	**95.2%**	**95.0%**
		Individual	**94.8%**	

In [11], M3 is trained with their crawled dataset. The authors treat the entire Humanizr and Demographer datasets as a test set and report the performance. The achieved accuracies are 80.7% for organization and 98.6% for individual accounts, whereas our model's performances on cross-validation are 90.4% and 98.5%.

Table 3.7: Performance details of the proposed model under k-fold cross-validation.

Dataset	Class	Precision	Recall	F1 Score	Accuracy
Humanizr(Full)	Organization	77.4%	71.8%	74.5	95.0%
	Individual	96.8%	97.6%	97.2	
Humanizr(Unbalanced)	Organization	70.6%	57.6%	63.3	92.5%
	Individual	94.8%	96.9%	95.8	
Humanizr(Balanced)	Organization	89.3%	85.6%	87.3	87.6%
	Individual	86.3%	89.6%	87.9	
Demographer(Full)	Organization	90.7%	90.4%	90.6	97.4%
	Individual	98.5%	98.5%	98.5	
Demographer(Unbalanced)	Organization	88.8%	87.8%	88.3	96.3%
	Individual	97.7%	97.9%	97.8	
Demographer(Balanced)	Organization	94.8%	95.2%	95	95.0%
	Individual	95.2%	94.8%	95	

Table 3.7 provides the proposed model's overall evaluation results since we provide additional evaluation metrics such as precision and F1-score besides metrics provided in baseline studies. The table presents the results of the proposed model on all versions of both Humanizr and Demographer datasets.

Table 3.8: Comparison of the proposed model using numeric, textual, and textual+numeric features on Demographer full dataset under k-fold cross-validation.

Features	Class	Precision	Recall	F1 Score	Accuracy
Only Numeric	Organization	52.0%	67.1%	58.6	90.0%
	Individual	96.0%	92.7%	94.3	
Only Textual	Organization	87.7%	**91.6%**	89.6	97.3%
	Individual	**98.7%**	98.1%	98.4	
Numeric+Textual	Organization	**90.7%**	90.4%	**90.6**	**97.4%**
	Individual	98.5%	**98.5%**	**98.5**	

We have also conducted experiments to determine the effects of the numerical and textual features on the classification performance separately. The desired model should produce a high accuracy value since the number of organizational accounts is much fewer than the individual accounts. However, more importantly, it should also have high precision values for both classes. As seen in Table 3.8, using only the numerical features produces 90% accuracy, but with a very low precision on organizational accounts. On the other hand, using only the textual features produces much higher accuracy and relatively high precision values for both classes. Moreover, adding the numerical features to the textual features seems to increase accuracy slightly. However, it has a significant effect on increasing the precision of the organizational account by 3%. Therefore, from these results, we can easily claim that combining the numeric and textual features gives the best model.

The results show that the metadata of the Twitter accounts contains crucial information about the classification of individual vs. corporation accounts. Textual features make the dataset more separable than the numerical features. Using embedding vectors of textual features instead of occurrence information of words for classification gives higher performance. According to our perspective, misclassification of an organization account is a more severe problem—the accuracy of our proposed model is 97.4%, which is the best among baselines.

Our evaluation highlights the robustness of the proposed model in distinguishing between individual and organizational accounts across different datasets and configurations. Notably, the combination of textual and numeric features proves to be a key contributor to the model's superior performance. As observed, precision for organizational accounts significantly improves when numeric features are included alongside textual ones. This improvement aligns with our expectation that metadata-driven insights can complement textual analysis to enhance classification accuracy. Additionally, the k-fold cross-validation ensures that the results are not biased by any specific data partition, thereby validating the generalizability of the proposed model.

Another important finding from our experiments is the balanced performance across multiple evaluation metrics. While other studies report strong metrics on specific datasets or classes, the proposed model consistently achieves high recall, precision,

F1-score, and accuracy for both individual and organizational accounts. This balanced performance is crucial for real-world applications, as it minimizes the risk of disproportionately misclassifying one class over the other. Our approach not only outperforms the baseline models but also underscores the value of integrating diverse feature types to achieve optimal classification results in tasks involving heterogeneous data.

3.3 Discussion

Classifying Twitter accounts as either individual or organizational entities presents a fundamental challenge with far-reaching implications for various applications, including user segmentation, targeted marketing, and social media analytics. The proposed approach in this chapter leverages account metadata, combining textual and numerical features, to achieve a robust classification framework. This method circumvents many limitations associated with text-based classification, such as language dependency and the noisy nature of user-generated content.

One of this study's significant findings is the effectiveness of metadata-based classification. The combination of textual features, processed through advanced subword tokenization techniques and numerical attributes such as follower count and engagement metrics, demonstrated superior performance compared to baseline methods. This result underscores the value of incorporating diverse metadata dimensions to enhance the separability of individual and organizational accounts. Moreover, using SentencePiece for subword segmentation effectively addressed challenges posed by multi-language and concatenated word inputs, making the approach scalable to a global dataset.

The experiments reveal that while numerical features alone can provide some insights into account classification, they are insufficient to achieve high precision and recall, particularly for organizational accounts. When combined with numerical data, textual features substantially improve classification performance, as evidenced by the precision, recall, and F1 scores reported across different dataset configurations. This finding highlights the importance of leveraging textual metadata, such as profile bios

and names, in conjunction with numerical attributes for a holistic classification approach.

Another noteworthy aspect of this study is the ability to generalize the model across diverse datasets, including balanced and unbalanced versions. The model's adaptability demonstrates its potential for deployment in real-world scenarios where the class distribution may be highly skewed. Additionally, the inclusion of protected accounts in the analysis, made possible by relying on publicly available metadata, further extends the applicability of this framework.

Despite these advancements, certain limitations warrant discussion. First, the reliance on metadata introduces potential vulnerabilities to account manipulations, such as fake or incomplete metadata fields. While the model performed well on the tested datasets, real-world applications may encounter adversarial behaviors aimed at evading classification. Future work could explore methods to detect and mitigate such behaviors, ensuring the robustness of the classification system.

Another limitation is the static nature of metadata, which may not capture dynamic changes in account behavior over time. Incorporating temporal features, such as activity patterns or engagement trends, could provide additional context to improve classification accuracy further. Furthermore, integrating network-based features, such as interaction graphs, may enhance the model's ability to differentiate between account types with similar metadata profiles.

The proposed approach demonstrates a significant step forward in metadata-based Twitter account classification. By reducing reliance on tweet content and focusing on profile attributes, this method offers a scalable, language-independent solution suitable for large-scale social media analytics. Future extensions of this work could explore the integration of additional metadata dimensions, real-time classification systems, and applications in broader social media contexts.

CHAPTER 4

TARGETED MARKETING ON SOCIAL MEDIA: UTILIZING TEXT ANALYSIS TO CREATE PERSONALIZED LANDING PAGES

Organizations attempt to utilize online tools to connect with potential customers by getting their attention, building attraction, and activating purchasing action. Commonly employed techniques include sending emails, the display of banner advertisements on pages, and enhancing visibility through search engine optimization. With the mainstream emergence of social media, various traditional marketing strategies have been adapted for application in the online sphere. One such strategy is viral marketing, a concept that is not novel but originates in pre-modern times when word of mouth was the only means of marketing goods [89]. An *active marketing* approach can be likened to this, rather than remaining in a physical store awaiting customers, actively seeking them out, and communicating the most salient features of products to them through their friends or lookalikes [90].

The three fundamental steps of online marketing are:

- **Acquisition:** The process of directing people to a specific website or landing page.
- **Conversion:** The act of persuading individuals to engage in specific desired actions.
- **Retention:** The task of fostering an ongoing relationship with visitors to a website, thereby increasing engagement and satisfaction levels.

Each phase of online marketing is intrinsically linked to the others, with success in one stage influencing the outcome of the subsequent stages. Despite the significance

of all three steps, many organizations focus their marketing expenditures on acquisition and retention while overlooking the conversion phase, specifically, the persuasiveness of a landing page, resulting in sub-optimal returns on advertising expenses [91].

The landing page serves as an initial interface between potential customers and the selling organization, wherein the visitors' attention needs to be captivated, and an attraction must be generated to persuade them to take action within a short time frame of their valuable consideration. Suppose they cannot perceive a pertinent connection between their requirements and the product's attributes at first glance. In that case, they will likely leave the page within seconds of eyeballing it.

The preliminary step in designing a marketing campaign is identifying the target audience. Although some social media platforms may facilitate the process through structured profile searches, identifying and defining the target audiences may be complex, or a website may need to offer detailed search & navigation tools so users can find their way. In such situations, an alternative approach is to search for shared topics of interest among customers, develop detailed target audience segment profiles, and develop targeted messaging containing relevant keywords emphasizing the specific problems and solutions the product meets. The use case we present addresses the challenge of identifying potential clients who are afflicted with high costs of drug prices, particularly those who have chronic diabetes and require life-saving long-term medication.

We have extended the model of the online marketing activity funnel presented by [91] by introducing a new step of target customer identification. The modified model, depicted in Figure 4.1, uses predefined keywords to filter the relevant messages, which aids in identifying genuine potentials who are more likely to become prospects. The primary objective is to persuade these potentials to visit the landing page by crafting a compelling message with a link to the page. Upon visiting the landing page, the potentials are then converted into prospects.

The terms potentials and prospects are often used interchangeably in marketing despite having subtle differences in meaning. Potentials refer to individuals or organizations who have the potential to become customers but have not taken any concrete

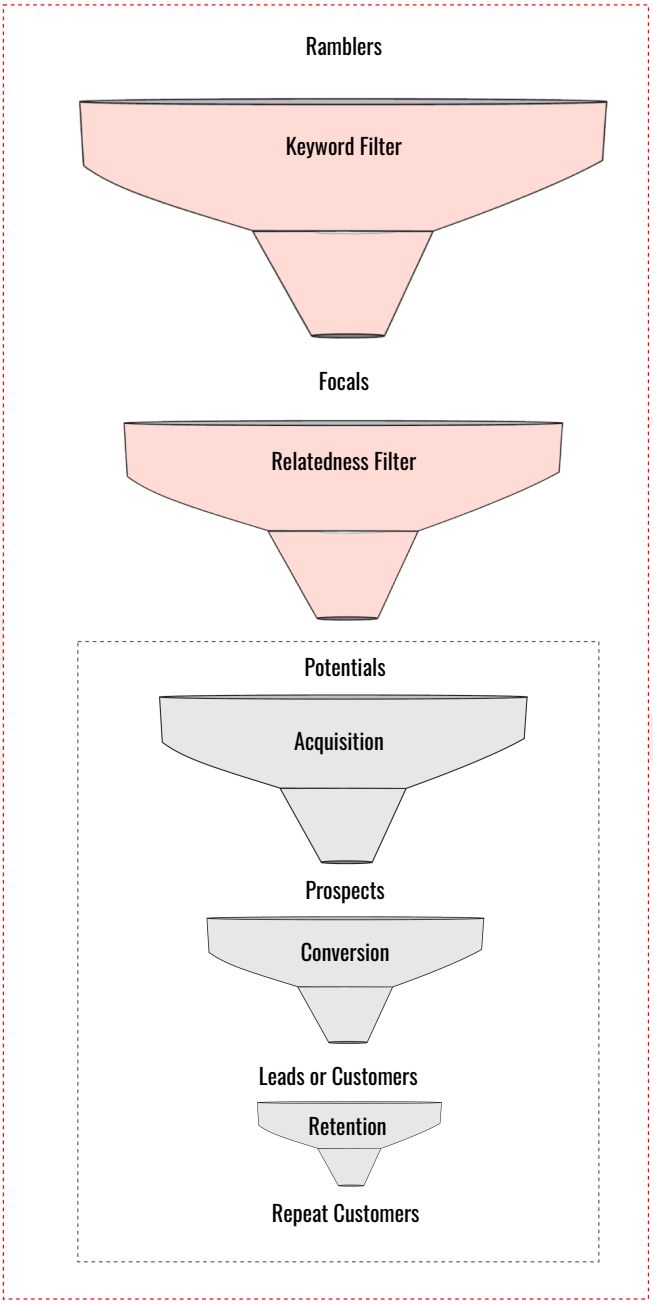


Figure 4.1: The extended version of the online marketing activity funnel.

action toward purchasing. Conversely, prospects refer to individuals or organizations who are actively considering the product or service as a viable solution to their needs or problems and have taken some action, such as visiting the landing page for more information. Potentials encompass a broader group of individuals who may not necessarily be actively considering the offering, while prospects constitute a more focused group considering the product or service. Nevertheless, both terms can be used depending on the context. Lastly, the study does not delve into retention as it requires a quantifiable metric for loyalty.

A product can have multiple functions to solve various issues. However, presenting all solutions on the landing page can result in a more cohesive design. Given that each client's priorities may differ, it is imperative to distinguish their individual needs and display a customer-focused page that aligns with their interests. This strategy would enhance the match between the customer's requirements and the product offering. For instance, patients unhappy with their health insurance provider's failure to cover the necessary medication should see a distinct campaign page from those dissatisfied with the additional costs. A discount card would address both issues even though the concerns are different. It is necessary to enumerate the problems the product addresses to identify the issues.

This study presents an innovative approach to improve online marketing activities for products or services that lack a clear definition of target customers. To the best of our knowledge, no prior study has explored creating a personalized landing page grounded in the priorities of potential clients. The proposed pipeline comprises three key steps;

- Identifying the keywords of interest to collect social media messages,
- Building a model to detect related messages without requiring extensive labeling,
- Generating a coherent landing page paragraph that utilizes the message context to deduce the customer's priorities.

An early version of this chapter has been published by [92] as a short paper. In this chapter, we have made several additions to the proposed pipeline described in the

early version. These include incorporating sentence transformers and RoBERTa in addition to BERT as a word embedding, training a classifier for relatedness on top of both BERT and RoBERTa models, adding BERTopic and RoBERTa as benchmarks for topic modeling in addition to BERT, GuidedLDA, and GloVe, and adding RoBERTa as a benchmark for sentence selection during paragraph generation in addition to BERT and GuidedLDA. We conduct experiments to validate these additions and show their superiority over existing methods. The code for the proposed pipeline can be accessed in the following repository [1].

The subsequent sections of this chapter are structured as follows. Section 4.1 focuses on data and methodology, providing a comprehensive overview of data collection processes, dataset summaries, and a step-by-step explanation of the proposed pipeline. Section 5.2 delves into the experiments and results, discussing each stage of the pipeline and its implications. Section 4.3 critically examines the limitations of utilizing large language models in this specific use case. Lastly, in Section 4.4, we draw conclusions based on our findings, summarizing key takeaways and suggesting avenues for future research and improvements in personalized landing page strategies.

4.1 Data and Methodology

The proposed approach comprises interconnected components that produce a customized, dynamically generated landing page. The pipeline steps are illustrated in Figure 6.4. The process begins with collecting tweets containing pre-determined keywords and filtering out unrelated ones. These relevant tweets are subsequently thematically modeled to extract priorities based on the codebook established by domain experts. Finally, topic probabilities are utilized to generate a coherent paragraph for the landing page that addresses the prospect's concerns.

[3]Çetinkaya et al. (2024). Targeted Marketing on Social Media: Utilizing Text Analysis to Create Personalized Landing Pages, *Source* [online]. Website https://github.com/tweetpie/personalized-landing-pages [accessed December 2024].

4.1.1 Data collection

It is crucial to carefully select the keywords utilized to get tweets from Twitter since keywords designated by humans are typically constrained and subject to personal bias. Instead, we locate communities relevant to the targeted issue on social media platforms. In this study, Facebook groups and pages where individuals with diabetes from the U.S. frequently interact are manually searched to pick posts from these communities. The details of the general corpus collected from Facebook about diabetes are shown in Table 4.1.

Table 4.1: The tabulated count of the posts procured from Facebook groups and pages concerning the diabetes.

Collected Posts	
Facebook Group #1	20,368
Facebook Group #2	6,696
Facebook Group #3	430
Facebook Page #1	2,574
Facebook Page #2	2,237
Facebook Page #3	1,506
Facebook Page #4	1,340
Facebook Page #5	513
Facebook Page #6	359
Total	36,023

To identify significant Twitter keywords, we employed the YAKE! (Yet Another Keyword Extraction) method [93] on the general corpus collected from Facebook. YAKE! is a language-independent, unsupervised technique that leverages statistical properties of the corpus [94]. Unlike traditional approaches requiring annotated corpora and extended training times, it uses statistical features from a single document for keyword extraction. It is adaptable to various languages, including minor ones, as it does not rely on named entity recognition or part-of-speech tagging. For evaluating phrase quality, criteria such as popularity, concordance, informativeness, and completeness are considered [95]. YAKE! provides popularity and completeness measures based on statistical properties, and further human labeling on the extracted keywords assesses

concordance and informativeness.

To generate a single document, we merge the posts and consider n-grams of up to three words in length. The top 2000 keywords are extracted for further evaluation by domain experts. The resulting list is subjected to human labeling, and the designated phrases are marked as valid and subsequently employed to search for relevant tweets. Table 4.2 shows the curated list of keywords.

Table 4.2: Some of the keywords extracted by YAKE! from collected Facebook posts. Is keyword column is the label provided by domain experts.

Phrase	YAKE! Score	Is Keyword
blood sugar level	4.50e-07	True
blood sugar	7.99e-07	True
type diabetes	1.72e-06	True
diabetes	3.33e-06	True
⋮	⋮	⋮
insulin	1.41e-05	True
type	1.49e-05	False
time	1.50e-05	False
doctor	1.51e-05	False
⋮	⋮	⋮
eat meal	6.37e-04	False
high glucose	6.38e-04	True
advance blood sugar	6.39e-04	True
time low	6.40e-04	False
⋮	⋮	⋮

In addition to the collected corpus, we developed another dataset called codebook to define the product's marketable aspects as seeds for the topic modeling. Before we proceed with topic modeling, we have preprocessed the tweets with the following steps;

- Emojis, URLs, user mentions, retweet sign (RT), multiple white spaces, etc., are pruned,
- Words are lowercased and lemmatized,

- Stopwords are excluded,
- Unigrams, bigrams, and trigrams are generated,
- n-grams having 0.01 min. document-frequency and 0.85 max. document-frequency is included in the vocabulary

Detected terms are then manually coded by domain experts for their categories. The process requires a comprehensive understanding of the data and identifying key themes and patterns. Determined categories, in turn, represent the product's commercializable features that we aim to promote. The potential customers' priorities align with eight distinct categories, namely *card, service, cost, insurance, health, medication, patients,* and *chronic*. Table 4.3 illustrates sample entries in the codebook and Table 4.4 represents the summary of the codebook.

Table 4.3: Some of the entries in the codebook created by domain experts of the sample use case.

Phrase	Synonym	Category
⋮	⋮	⋮
coupon cards	coupon	card
discount provider	coupon	card
⋮	⋮	⋮
drugs cost	drug price	cost
healthcare cost	health costs	cost
⋮	⋮	⋮
medicare	insurance coverage	insurance
employee plans	insurance coverage	insurance
⋮	⋮	⋮
humalog insulin	diabetes medication	medication
januvia	diabetes medication	medication
⋮	⋮	⋮

To summarize, the primary objective of the first corpus, messages from Facebook groups and pages, is to identify important keywords that help close monitoring of the

Table 4.4: Codebook summary.

	Phrase #	Synonym #	Avg. length
card	137	13	15.0
service	266	23	23.0
cost	231	27	13.9
insurance	101	15	15.1
health	101	17	10.4
medication	400	25	13.1
patients	53	8	15.5
chronic	30	5	9.0

Twitter platform. The second corpus, which we obtain from Twitter, comprises tweets containing the keywords identified in the first dataset. The last dataset is the manually built codebook used while applying topic modeling to the target tweet. Marketable aspects are derived from the codebook.

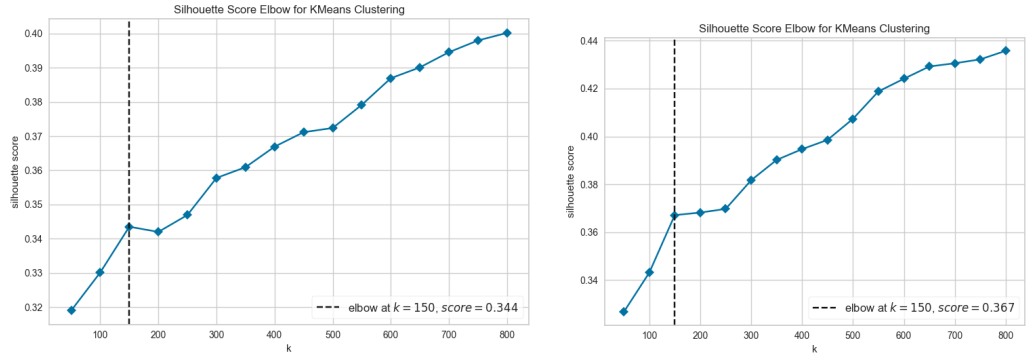
4.1.2 Classifying Related Tweets

We acknowledge that presuming that tweets containing particular keywords are related to the intended domain represents a considerable bias. Therefore, the collected tweets must be carefully curated by assessing their context to eliminate irrelevant content, thereby retaining only the pertinent material. It ensures the corpus is sufficiently extensive, providing insight into the most commonly utilized phrases or words.

Given the labor-intensive nature of collecting and labeling 70,972 tweets, our research employs an efficient strategy centered around leveraging pre-trained language models renowned for effectively embedding contextual information. We extract meaningful features from tweets using models such as BERT, RoBERTa, and sentence transformers, generating 768-dimensional sentence embeddings vectors. After applying Principal Component Analysis (PCA), visual assessments reveal discernible clusters of semantically similar sentences, as depicted in Figure 4.2. Applying the K-Means algorithm, coupled with a grid search for optimal cluster determination, yields the best silhouette score for $n = 150$ for both approaches, as demonstrated in Figure 4.3,

indicating a high level of compactness within the identified clusters.

Figure 4.2: The visualization of the PCA applied to the tweets' BERT embeddings in 150 clusters, viewed from different angles. Best viewed in color.



(a) K-Means using BERT embeddings

(b) K-Means using RoBERTa embeddings

Figure 4.3: The silhouette scores for clustering all collected tweets utilizing the K-Means algorithm while varying the number of clusters k.

The labeling process for our dataset involves leveraging the semantic proximity of samples within clusters, mitigating the labor-intensive task of individually labeling each tweet. Instead, we label the top 100 tweets closest to the centroids in each cluster, attributing the cluster's label to all members. This strategy allows for the efficient labeling of a subset of the data, ensuring that the labeled tweets accurately represent their respective clusters. The summary of the weakly-labeled dataset, obtained through K-Means clustering on BERT, RoBERTa, and sentence transformers models, is presented in Table 4.5. Notably, the unrelated tweets in the dataset are 2 to 3 times the related ones. The weakly labeled dataset is used to train a classifier using pre-trained language models and K-Means voting.

For the language-model-based classifiers, the models are fine-tuned by adjusting their weights and biases to align with categorizing tweets into relevant and irrelevant categories, as displayed in Figure 4.4. The architecture includes a fully connected layer atop BERT/RoBERTa and sentence-transformers, a normalization layer with tanh activation, and a dropout layer. A second fully connected layer is then applied to refine the features extracted from the tweet. We first pre-trained the fully connected layers

Table 4.5: Dataset summary which is labeled using K-Means clustering on top of BERT, RoBERTa embedding and corresponding sentence-transformer models displaying related/unrelated clusters.

	BERT		RoBERTa		S-BERT		S-RoBERTa	
	Related	Unrelated	Related	Unrelated	Related	Unrelated	Related	Unrelated
Cluster #	46	104	40	110	49	101	42	108
Tweets #	25,261	45,711	18,772	52,200	24,312	46,660	20,271	50,701
Min cluster	169	19	151	58	284	211	237	198
Max cluster	1077	996	700	997	788	770	715	702
Avg. cluster	549.2	439.5	469.3	474.6	496.2	462	482.6	469.5
Min text	19	4	15	4	4	6	6	4
Max text	154	158	154	158	158	158	158	158
Avg. text	128	115.8	126.1	118	128.3	117.4	127.2	118
Min word #	2	1	2	1	1	1	1	1
Max word #	32	32	31	32	32	31	32	32
Avg. word #	18.5	16.8	17.8	17.2	17.4	15.8	17.8	15.9

Table 4.6: Cosine similarities between 'sweet' words' BERT embeddings in different contexts.

Sentence 1	Sentence 2	Cos. Sim.
Eating lots of **sweet** worsens joint pain due to inflammation.	Kind **sweet** cat on the sofa.	0.46
People with diabetes can eat **sweet** with a healthful meal plan.		0.84

to gradually learn higher-level features and improve the model's accuracy in tweet classification.

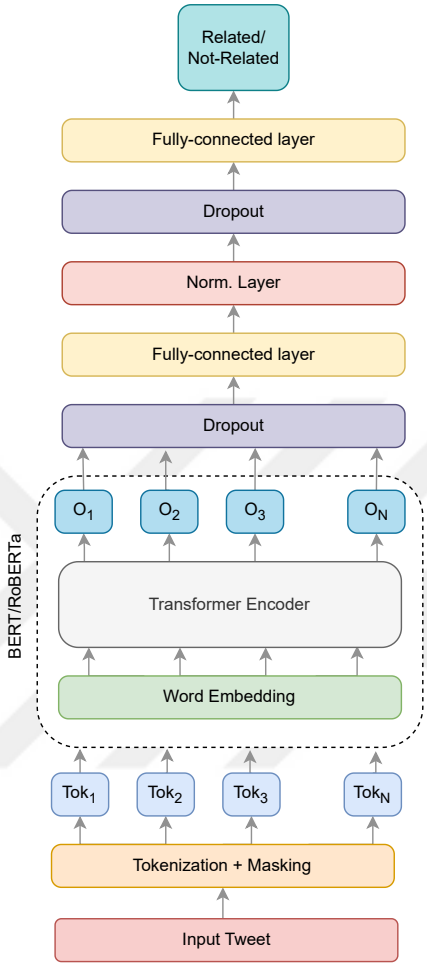


Figure 4.4: Pre-trained language model based classifiers' model diagram.

The second classifier in our study employs the K-Means clustering model used for dataset preparation, with the majority voting to determine whether given tweets are relevant. This classifier utilizes separate BERT and RoBERTa representations for each cluster and domain expert-provided labels. For a given tweet, the closest n clusters are determined based on the embeddings, and the labels of each cluster are examined. The valid label is assigned if the majority of the labels in the cluster indicate the tweet's relevance.

4.1.3 Topic modeling for user's priorities

Thematically modeled topics are employed to extract the priorities from the tweets. GuidedLDA [45], an extension of LDA [96], overcomes limitations in explaining only superficial aspects of a corpus by incorporating "guide seeds". The seed words enhance coherence and interpretability. Guided BERTopic [97] utilizes BERT embeddings for topic modeling, grouping similar documents. It converts seeded topics into embeddings, assigning labels based on cosine similarity. UMAP [98] is used for dimensionality reduction in BERTopic since it can capture both the local and global high-dimensional space in lower dimensions. A multiplier is applied to seed topic words, increasing IDF values and enhancing their appearance in topics—BERTopic handles large datasets, showing competitive performance [46].

We aim to leverage these topics to order paragraphs based on user interests, employing unsupervised topic modeling to detect document themes. Alongside GuidedLDA and BERTopic, we explore BERT, RoBERTa, and GloVe embeddings for clustering and topic extraction since the similarity in the context of the documents makes deriving the latent themes possible [99]. We create cluster centers using embeddings of seed words related to topics, evaluating cosine similarity distances of sentence embeddings for topic selection. The resulting topic distribution orders paragraphs by relevance to user interests, facilitating the creation of a personalized landing page.

As mentioned above, domain experts define the main marketable aspects of the product as a codebook. Later, we prepare a set of generic sentences as advertising copies for each. There are eight particular categories, and Table 4.7 gives the details. The user's message may be prone to multiple categories with different weights corresponding to the user's priorities. For instance, the tweet *"Skyrocketing drug prices are hurting patients and costing taxpayers. Millions of Americans do not take their medications as prescribed due to the price."* has complaints about costs, medication, and patients.

Table 4.7: The codebook-derived marketable aspects of the product and the accompanying general sentences for each aspect.

Category	Sentences About
card	An explication of the discount card and its underlying purpose
service	The scope of the Brand's service and its accessibility to the customers
cost	Concerns regarding the excessive payment for prescriptions
insurance	The limitations and disadvantages of the traditional insurance model
health	The pivotal role of health and its impact on an individual's life
medication	The significance and value of medication in treating ailments
patients	The target population who can benefit from the Brand's products
chronic	Elucidation on chronic illnesses and their implications.

4.1.4 Generating a coherent landing page paragraph

Utilizing extracted topic probabilities, we generate concise landing page paragraphs addressing prospect concerns validated through human evaluation. Recognizing the need to convince users to be consciously aware of the promotional nature of the landing page, our strategy aims for immediate resonance with their needs. Starting with the topic about the prospect's highest priority and progressing to the least, the landing page paragraph is limited to a maximum of five sentences to enhance user readability. For cases where the user's message encompasses more than five topics, top ones are prioritized. Each topic is associated with generic sentences discussing various aspects, and the most relevant sentences from selected topics form a coherent paragraph. To gauge relatedness, sentences, and the user's message are converted into vectors using the trained GuidedLDA model, with cosine similarity determining the adoption of the most similar sentence.

Considering the coherency in the paragraph, we apply two different approaches;

1. Each topic sentence is selected, independently calculating the similarity between candidate sentences and the source message. The most semantically similar sentence is selected for every sentence.

2. The first sentence in the paragraph is selected by looking at the similarity be-

tween the highest priority topic candidates and the source message. Upcoming sentences are selected by comparing the first selected sentences' and the candidate sentences' semantics.

For the running example in the previous section that identifies the topic, the selected sentences among candidates for the paragraph using both approaches are displayed in Table 4.8. Using only the source tweet while selecting the following sentences will produce a paragraph highly related to the source sentence. However, the incoherent behavior in the paragraph can be easily sensed. Similarly, using only the previous sentence solves the coherency problem, but the latter topics are slightly becoming disconnected from the source tweet.

4.2 Experiments

This section presents the experiments conducted on each phase of the proposed pipeline and discusses the results.

4.2.1 Related Tweet Classification Experiments

To validate the relatedness model's performance, we randomly collected 174 tweets containing determined keywords and labeled them, with 82 being related and 92 unrelated. We convert tweets into word embeddings using the same procedure as we develop the relatedness model and calculate the cosine similarities for each cluster center. The most similar n clusters vote on the tweet based on their label and assign the majority's decision. In order to determine n value, we apply grid search on a separate set, which gives 7 as the best f-score.

Predicting unrelated tweets as related is a more serious problem since it will annoy people not interested in the product. We adjusted the model to get a more decisive majority where 5 out of 7 should vote as related. We implement two variations of majority voting, namely, simple majority voting and strong majority voting, the latter utilizing a threshold to minimize the likelihood of false positives. Similarly, we employ two logit thresholds, 0.5 and 0.95, respectively, for the pre-trained language

Table 4.8: Generated paragraph for the tweet *"Skyrocketing drug prices are hurting patients and costing taxpayers. Millions of Americans don't take their medications as prescribed due to the price."* using two different approaches.

Topic	Selected Sentences
	Only source tweet
cost	The cost of a prescription may differ by more than $100 between pharmacies across the street from each other!
medication	Taking the right medications at the right time is not only important financially—it is also essential to health and wellbeing.
patients	[BRAND] works for everyone: insured, uninsured, Medicare recipients.
card	[BRAND] is working to create a healthier wealthier America.
	Only previous sentence
cost	The cost of a prescription may differ by more than $100 between pharmacies across the street from each other!
medication	Taking medicine as prescribed or medication adherence is important.
patients	With the rising price of prescriptions and the increase in high deductible health plans and high copays, there's no limit to who [BRAND] can help.
card	[BRAND] will help you pay less than the cash price for prescription.
	Static landing page paragraph
	Say goodbye to soaring drug prices and hello to accessible, budget-friendly healthcare. With our card, you'll unlock exclusive discounts at pharmacies nationwide, ensuring you never have to compromise your health due to cost. Join the thousands who have already embraced smart, cost-effective healthcare – subscribe [BRAND] today and prioritize your well-being without breaking the bank!

model-based classifiers.

Table 4.9: Relatedness classification performance results of the models with different threshold values.

Model	Θ	Recall	Precision	Accuracy	F-1
K-Means (BERT)	4/7	0.77	0.76	0.78	0.76
	5/7	0.41	0.85	0.69	0.55
K-Means (RoBERTa)	4/7	0.89	0.71	0.78	0.79
	5/7	0.79	0.73	0.76	0.76
K-Means (S-BERT)	4/7	0.85	0.63	0.70	0.73
	5/7	0.73	0.65	0.69	0.69
K-Means (S-RoBERTa)	4/7	0.83	0.64	0.70	0.72
	5/7	0.68	0.65	0.68	0.67
LM-based (BERT)	0.5	**0.95**	0.77	0.85	0.85
	0.95	0.90	0.85	**0.88**	**0.88**
LM-based (RoBERTa)	0.5	0.82	0.77	0.80	0.79
	0.95	0.71	0.83	0.79	0.76
LM-based (S-BERT)	0.5	0.92	0.77	0.83	0.84
	0.95	0.87	**0.88**	**0.88**	0.87
LM-based (S-RoBERTa)	0.5	0.84	0.72	0.77	0.77
	0.95	0.72	0.84	0.80	0.78

The tabular data in Table 4.9 illustrates that the K-Means model, utilizing BERT embeddings from sentence transformers, achieves high precision scores when employing a threshold of 5/7. In contrast, the LM-based model, with the sentence transformer RoBERTa embeddings, attained the highest recall score (0.95) at a threshold value of 0.5 while maintaining a commendable precision score (0.77), resulting in the best F-1 score (0.85). When Θ is set to 0.95, the LM-based models demonstrate precision performance comparable to the K-Means model using the same embedding method yet maintaining higher recall scores. This study's findings indicate that LM-based classifier models outperform K-Means models across precision, recall, accuracy, and F-1 score. However, no discernible pattern emerges regarding the embedding method. Even when weakly labeled data is utilized through clustering, LM-based models integrated with clustering models exhibit superior generalization capabilities.

4.2.2 User Priority Extraction Experiments

This study employs guided topic modeling to incorporate domain expertise in detecting user priorities. To assess the models, three participants ranked the top five categories for 20 tweets in relevance, validating the topic model. Fleiss' Kappa [100] is calculated to measure annotator agreement, yielding a result of 0.643. Strong agreement is reached for the remaining 15 tweets through a one-out technique to address disagreements, resulting in a Kappa score of 0.824. The models, including BERT, RoBERTa, GloVe, GuidedLDA-based, and BERTopic-based, extract topics with probabilities, and their overlaps are benchmarked against user-assigned ranks. Table 4.10 presents the model comparisons based on exact matches, +/-1, +/-2 rank orders, and unordered matches. Consistent with prior research [42], BERT exhibits topic predictions close to GuidedLDA, particularly without considering order, followed by GloVe. BERTopic performs similarly to BERT, attributed to its reliance on seed words and their embeddings. However, GuidedLDA excels in prioritizing predicted topics, as observed in the prediction of tweet topics.

Table 4.10: The ranked scores of the top-five topics predicted for a given tweet based on the user-provided ranks, which are sorted by their rank-oriented scores with different tolerance values.

Model	Exact	+/-1	+/-2	Any
BERT	0.6	1.6	2.7	3.1
RoBERTa	1.1	1.9	3.0	3.3
GloVE	0.4	1.3	2.0	2.9
BERTopic	0.7	1.8	2.5	2.9
GuidedLDA	**2.7**	**3.3**	**3.5**	**3.7**

The GuidedLDA model, furnished with seed words representing the corpus and incorporating two blank topics for comprehensive coverage, reveals identified topics in Table 4.11. The model autonomously generates the last two topics, introducing dimensions related to hope, fighting against fatal diseases, and focusing on a healthy life, aligning with the predefined codebook.

Each topic is weighted, with the sum of weights equating to 1, and contains a set of words with normalized weights totaling 1. When a document is processed, the

significance of each phrase for the corresponding topic is calculated, resulting in a final topic probability distribution. These calculated probabilities serve as message priorities during paragraph generation.

The model results in the topic distribution for the example about "skyrocketing drug prices" covered in this section to 68% cost, 20% medication, 10% patients, and 1% card. The results are highly consistent, which detects the priorities of the message as expected.

Table 4.11: The identified topics by GuidedLDA are exhibited using given seeds, displaying the top 7 n-grams.

Topic	Words
card	gift, card, holiday, gift card, credit card, coupon
service	healthcare, ai, technology, problem, security, solution, medtech
cost	price, cost, save, saving, pricing, discount, money
insurance	insurance, plan, medicare, health insurance, coverage, pay, medicaid
health	health, health care, care, digital, mental, aging, clinic
medication	medication, med, diaetes, flu, vaccine, insulin, pill
patients	patient, pharmacist, physician, cancer, medical, therapy, clinical
chronic	disease, chronic, disability, dementia, risk, chronic pain
new topic-1	hope, family, member, dream, impact, cancer, fight
new topic-2	love, holiday, stay, healthy, book, support, eat

4.2.3 Landing Page Paragraph Generation Experiments

Propaganda phrases for each category are pre-prepared, and the paragraph generator constructs paragraphs by sequentially selecting one sentence from each category, following the order determined by the topic model. Two techniques for sentence selection, employing different models, are compared. The first approach prioritizes generating paragraphs closely related to the target tweet, calculating cosine similarity between the tweet and candidate sentences' representations for each topic. The second approach emphasizes coherence by evaluating similarity to the previously selected sentence, utilizing embeddings from BERT and GuidedLDA models.

The cosine similarity between consecutive sentences is calculated in the evaluation using 120 tweets, generating 600 sentences. Table 4.12 reveals that BERT, employing distance to the previous sentence, yields 125 divergent sentences, undermining semantic integrity. While RoBERTa outperforms GuidedLDA in divergent sentences with 139 out of 600, it has fewer cosine similarity values due to its self-calculated representations. Consequently, BERT is selected for specific categories identified by the GuidedLDA topic model.

Table 4.12: The divergence and cosine similarity metrics of the selected next-sentence embedding model with respect to the average of the previous sentence.

Model	Δ to	Divergent Sentences	Cos. Sim.
BERT	source tweet	35.3%	0.789
	previous sentence	**20.8%**	**0.834**
RoBERTa	source tweet	39.3%	0.622
	previous sentence	23.1%	0.742
GuidedLDA	source tweet	45.2%	0.721
	previous sentence	27.6%	0.795

4.2.4 Human evaluation experiments

The pipeline's performance evaluation involves collecting tweets from Twitter through predefined keywords, filtering relevant tweets via the relatedness model, extracting topic models using GuidedLDA, and generating paragraphs by selecting sentences based on BERT embeddings. Ten tweets and two pages for each are presented to five participants, who, assuming the role of the tweet owner, visited both a static and personalized landing page, rating their attention and persuasion from 1 to 5. Table 4.13 indicates that 84% of participants found personalized landing pages attention-grabbing, compared to 56% for static landing pages. The personalized landing page receives an average score of 4.28 for addressing the issue. Fleiss' Kappa scores of 0.619 for attention and 0.667 for addressing the need indicate substantial agreement among raters.

In order to prepare the survey without bias, displayed tweets are chosen to consist

Table 4.13: The results of the human evaluation that compares the effectiveness of the static landing page and the personalized landing page proposed in this study.

Strategy	Attention Grab	Addressing the Need
Static Landing Page	56%	3.36
Personalized Landing Page	**84%**	**4.28**

of different topics. Since the static landing page mainly mentions the solution for additional costs, the participants thought that was related to the tweet. Excluding the tweets complaining about costs results in a 32% attention grab and a 2.85 addressing need score for the static landing page. The results show that the proposed pipeline performs better in attracting attention and providing solutions to the needs.

4.3 Employing Large Language Models for Personalized Landing Pages

In the field of NLP, LLMs have proven to be highly effective in solving a wide range of language tasks [101], showcasing proficiency across various applications. Notably, their capabilities extend beyond general-purpose assignments, encompassing the capacity to craft personalized landing pages tailored to specific needs and preferences. However, despite their significant contributions, it is essential to examine potential challenges associated with LLMs in this particular application. This section will explore and critique issues such as hallucination problems, template rule violations, and over-specificity, providing instances where LLMs may unintentionally deviate from the intended personalized design and message for individual tweet owners. Despite creating diverse prompts, ranging from minimal to comprehensive information, by incorporating company descriptions, tweet topics, and predefined sentences, we encountered issues such as hallucination problems, template rule violations, and over-specificity at every level of provided information. These observations underscore the imperative for continued refinement and cautious implementation of LLMs in generating personalized landing pages.

We conduct our experiments with prominent language models, including GPT-3.5 [102], GPT-4 [103], and Mistral [104]. Chat-GPT-3.5 and Mistral exhibit similar

limitations, particularly concerning hallucination problems and template rule violations. However, ChatGPT-4 demonstrates improvements, providing better outputs in discerning hallucination and template commands. Notably, ChatGPT-4 still needs to address over-specificity despite these enhancements.

The hallucination problem in LLMs is manifested when these models, even when guided by a knowledge base, may generate incorrect information [105]. LLMs can fabricate entirely inaccurate details if a tweet references content that is not in the provided knowledge base. To mitigate this issue, approaches such as few-shot learning [106], and Chain-of-Thought Prompting [107] can enhance the model's ability to avoid hallucinations. However, the efficacy of these methods remains uncertain, underscoring the necessity of thoroughly scrutinizing every output.

In the proposed model, the generation of landing pages is circumscribed by predefined sentences. This constraint is a safeguard, reducing the imperative to scrutinize each output while preserving a personalized touch. Below, you can find two examples illustrating the challenges associated with hallucinations. In the first example, the LLM creates an output containing the information that the product is designed for the underinsured population, which needs to be corrected when cross-referenced with the knowledge base provided. In the second example, the LLM states, "50% of the population benefitting from our money-saving tools," whereas the correct information in the knowledge base is "over 50% of the population can benefit from our prescription drug savings card."

Additionally, we observed that, at times, the LLM may alter the numerical values in the knowledge base. For instance, our knowledge base indicates one can save up to 70% of the money. However, inquiring about a 90% discount, the LLM might mention that one can save 90%. These instances underscore the importance of careful verification to ensure the accuracy of the generated content.

Tweet

There has to be a prescription savings card that I can use instead of my insurance.

Knowledge Base

... individuals can achieve cost savings, given the prevalence of high deductible health plans, elevated copayments, and a substantial portion of the population being underinsured ...

Landing Page Content

... our prescription savings card is designed for the underinsured population ...

Figure 4.5: Hallucination Example - Proceed from Tweet Source

Tweet

Struggling with pricey meds, can't afford without a solid discount on my insulin drugs.

Knowledge Base

... over 50% of the population can benefit from our prescription drug savings card ...

Landing Page Content

... With over 50% of the population benefitting from our money-saving tools ...

Figure 4.6: Hallucination Example - Proceed From Knowledge Base

The challenge of LLMs breaking template rules becomes even more pronounced when crafting personalized landing pages for individual tweets. While adept at generating coherent text, LLMs may struggle to adhere to predefined templates, strictly

disrupting the intended structure. This can result in visually inconsistent and less engaging user experiences, challenging the effectiveness of personalized landing pages in conveying a compelling and unified message tailored to candidate customers. For instance, consider a prompt directing an LLM to generate a landing page with a specific length: "Generate a 5-sentence length landing page." Despite such explicit instructions, the LLM may produce an output exceeding the specified length, potentially containing 8-9 sentences. This highlights a notable aspect of LLM behavior, where even when directed to generate content of a specific length, the model may exhibit a degree of unpredictability in adhering strictly to the given constraints, introducing challenges in maintaining precise control over the output.



Figure 4.7: Template Description in the Prompt

In the context of creating personalized landing pages, Language Models (LLMs) tend to over-specify. This tendency manifests when LLMs excessively tailor the content to individual tweets, emphasizing minute details to an extent that may neglect the broader context or diverse audience. Striking an optimal balance between personalized messaging and broader appeal is essential to ensure that the tailored landing pages effectively capture the essence of the customer's tweet while resonating with a more extensive range of audiences. Over-specificity may inadvertently create a sense of individual targeting, potentially causing concern for the tweet owner. Therefore, being mindful of this delicate balance ensures that our personalized messaging is a powerful tool that connects on an individual level without sacrificing the broader appeal for a diverse range of readers.

Tweet

I spend a significant amount on my Hutchinson-Gilford Progeria Syndrome medication and am seeking a discount specifically for this condition. I don't require assistance with anything else.

Prompt

[Knowledge Base], [Tweet], Create a 5-sentence length landing page paragraph for me targeting this tweet. It should be designed as a general landing page, strategically aligning with the content of the tweet without directly responding to its specific statements.

Landing Page Content

... We understand the challenges patients face, whether it's Hutchinson-Gilford Progeria Syndrome or any other medical condition ...

Figure 4.8: Over-Specificity Example - Proceed From Tweet Source

While addressing the limitations associated with Language Models (LLMs) in creating personalized landing pages is feasible, achieving faultless output every time is not guaranteed, necessitating vigilant verification. Overcoming these challenges involves prompt engineering, incorporating strategies such as utilizing chat memories, multiple prompt querying, and employing more complex prompts. However, it is crucial to acknowledge that these enhancements come at an increased cost, further complicating an already resource-intensive process when employing LLMs for each specific task. Therefore, despite the potential advancements, a judicious balance between addressing limitations and managing associated costs must be struck in the practical integration of LLMs into personalized response approaches. It is essential to note that even if the additional cost associated with avoiding these limitations by creating a more complex prompt or pipeline is accepted, the inherent nature of LLMs as a black box system makes outcomes unpredictable.

4.4 Discussion

This chapter addresses the critical challenge of enhancing the effectiveness of online marketing by leveraging social media data to create personalized landing pages. The proposed framework demonstrates the potential of integrating text analysis techniques with marketing strategies to generate coherent and engaging content tailored to specific user segments. The findings underscore the value of using advanced text processing methodologies to bridge the gap between audience understanding and practical marketing applications.

A key contribution of this chapter is the introduction of a comprehensive pipeline that combines user segmentation, topic modeling, and personalized paragraph generation. Topic modeling techniques, such as BERTopic, to identify latent themes in social media conversations enable the extraction of actionable insights that directly inform content creation. As demonstrated through experiments, the ability to capture user priorities and preferences highlights the framework's relevance in tailoring marketing messages to align with audience needs.

One of the most notable results is the significant improvement in user engagement and conversion rates achieved through personalized landing pages compared to static ones. The generated content resonates more effectively with target audiences by incorporating contextual relevance and thematic coherence. This finding emphasizes the importance of customization in digital marketing and validates the efficacy of AI-driven approaches in enhancing content relevance.

Integrating large language models (LLMs) for paragraph generation further strengthens the framework's capability to produce high-quality, human-like text. Using context-aware sentence arrangement ensures that the output is coherent and persuasive, addressing a common limitation in automated text generation. The results of human evaluation experiments, which favorably compare personalized content to non-personalized alternatives, highlight the practical applicability of the proposed methodology.

Despite these advancements, certain limitations must be acknowledged. First, the reliance on predefined keywords for topic modeling introduces a potential dependency on the quality and comprehensiveness of the seed terms. While this approach

provides a structured means of guiding topic discovery, it may exclude emerging or less apparent themes that are equally relevant to the target audience. Future research could explore adaptive methods for keyword selection or unsupervised alternatives to mitigate this limitation.

Second, the framework's success depends on the availability and quality of social media data. While platforms like Twitter and Facebook offer rich data sources, privacy regulations and API limitations may restrict access, potentially affecting the approach's scalability. Developing methods to work with anonymized or aggregated data while maintaining personalization could enhance the proposed solution's robustness.

Another consideration is the computational complexity of integrating advanced text analysis and language models. While the framework demonstrates strong performance, its implementation at scale may require substantial computational resources, particularly for real-time personalization. Optimizing model efficiency and exploring lightweight alternatives for deployment could address this challenge.

In conclusion, this chapter demonstrates the transformative potential of AI-driven methodologies in targeted marketing. The proposed framework bridges the gap between understanding audience preferences and delivering tailored content by combining text analysis with personalization strategies. Future work could focus on expanding the framework to incorporate additional data modalities, such as multimedia content, and exploring its application in diverse marketing contexts. The insights and methods presented here lay a strong foundation for advancing personalization in digital marketing.

CHAPTER 5

MASKING THE BIAS: FROM ECHO CHAMBERS TO LARGE SCALE ASPECT-BASED SENTIMENT ANALYSIS

Aspect-based sentiment analysis (ABSA) is a subset of sentiment analysis that not only discerns the sentiment within a given text but also links it to specific entities or aspects. This precision enhances understanding of stance, enabling scholars to derive actionable insights by pinpointing the sentiments on specific entities or issues. As a result, ABSA has become a critical tool in data analytics and AI applications [108].

While current ABSA methodologies have proven valuable, they require reevaluation to effectively handle the complexities and vast diversity present within large real-world datasets. One of the primary challenges of ABSA is training set bias, a phenomenon where the data used to train an algorithm fails to comprehensively represent the diversity of contexts and opinions in the full spectrum of the target populations. This bias can skew the algorithm's performance, often to the detriment of minority stances on less represented views, thereby compromising ABSA systems' efficacy and fairness.

In our model, we simply *mask* the terms corresponding to the aspects we seek to analyze. This technique allows our model to focus on the context alone to infer the sentiment directed towards the masked aspect without being biased by the aspect term itself. By doing so, we refine the model's capacity to more precisely gauge sentiments towards both previously seen and unseen (new) aspects, thereby addressing one of the critical limitations of current ABSA models.

The construction of labeled datasets for NLP tasks presents challenges due to the costly nature of labor-intensive manual coding. The high costs associated with ABSA

dataset development result in publicly available datasets being scarce. ABSA differs from traditional sentiment analysis by marking specific aspects in every sentence and labeling the sentiment corresponding to each aspect. This process necessitates understanding the domain-dependent key aspects and the contextual interplay between aspects and intended sentiments, which cannot be easily automated. This requirement for manual labeling significantly contributes to the high costs associated with ABSA dataset development and the lack of publicly available datasets. The limited number and size of available datasets pose challenges for reliable training, testing, and comparing models, especially transformer-based ones with millions of parameters. Researchers currently try to determine a small model's accuracy by averaging outcomes across multiple runs. However, larger datasets would facilitate more accurate evaluations [109].

Data analysis on social media platforms like X (formerly Twitter) is especially relevant due to such platforms' pervasive roles in enabling viral information cascades that impact public discourse. Recent studies [110],[21],[111] show that social media users pay attention to opinions they agree with. It indicates a strong correlation between biases in the content people both produce and consume. In other words, *echo chambers* reinforcing certain biased stances and a tailored media experience, that eliminates opposing viewpoints and differing voices, are very real on social media. Our approach employs community detection algorithms on retweet (or share) networks to reveal camps and their more homogeneous sub-communities. This method allows us to use partisan 'barrier-bound' users within a camp to rapidly and cost-effectively produce stance annotated weakly-labeled ABSA datasets.

We summarize our key contributions as follows:

- We foster large-scale, cost-effective, weakly labeled ABSA datasets by harnessing homogeneous stances of partisan users in social media communities.
- We unveil MaskedABSA, a model that enhances aspect-based sentiment analysis by masking aspect terms in sentences. This improves accuracy in gauging sentiments toward both known and previously unseen aspects.
- We evaluate the MaskedABSA model's performance under imbalanced per-

spective conditions, particularly focusing on accuracies for detecting underrepresented minority stances to bolster the fairness of ABSA systems.

- We make experimental datasets and models publicly available via GitHub at tweetpie/masked-absa.

The remainder of this chapter is structured as follows: Section 5.1 details the procedures involved in preparing the MaskedABSA and the weakly labeled dataset. Section 5.2 presents various experiments and evaluations. Section 5.3 offers concluding remarks and future works.

5.1 Methodology

In this section, we outline the approach and techniques used to address ABSA's challenges, including data scarcity, model biases, and the need for context-aware sentiment classification. Our methodology integrates the MaskedABSA model and weakly labeled datasets to enhance ABSA tasks' performance, scalability, and robustness.

The proposed methodology begins with creating weakly labeled datasets that leverage social network structures and community detection algorithms. By analyzing user interactions and discourse patterns, we generate large-scale datasets that can effectively capture the nuanced sentiment orientations of different communities. These datasets are refined through automated machine-learning techniques and expert reviews to ensure high relevance and quality.

5.1.1 Masked Aspect Sentiment Classification (MASC)

ABSA is a task that provides a polarity of each aspect term in the sentence. Therefore, Zhang et al. [112] define it as a comprehensive opinion summary at the aspect level. The process of aspect-based sentiment analysis comprises two distinct components: Aspect Term Extraction (ATE) and Aspect Term Sentiment Classification (ATSC). First, in the Aspect Term Extraction stage, the model identifies and extracts aspect terms from a provided sentence. Next, the model predicts the stance associated with

identified aspect terms in the Aspect Sentiment classification phase. As a result, the model returns aspect term and stance pairs. This task is illustrated in Figure 5.1. We formally define aspect-based sentiment analysis process as follows. Consider a

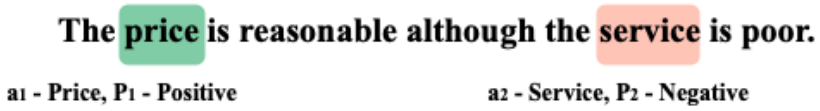


Figure 5.1: Example of aspect-based sentiment annotation.

sentence S_i expressed as:

$$S_i = \{w_1, w_2, w_3, \ldots, w_n\} \tag{5.1}$$

where each w_j represents a word in the sentence. Within the sentence S_i, we identify multiple aspect terms denoted by A_i:

$$A_i = \{a_1, a_2, \ldots, a_k\},\ where\ k <= n \tag{5.2}$$

The process of detecting and extracting these aspect terms is termed Aspect Term Extraction (ATE). Subsequently, each aspect term is associated with a polarity P_i:

$$P_i = \{p_1, p_2, \ldots, p_k\},\ where\ k <= n \tag{5.3}$$

Each polarity p_x is categorized as one of the following values:

$$p_x = \{pos, neg, neutral, conflict, none\} \tag{5.4}$$

For the sample sentence in Figure 5.1, the model extracts aspects a_1 (price) and a_2 (service). It then determines the polarities as p_1 (positive) and p_2 (negative), respectively. As a result, the model outputs pairs such as (a_1, p_1) and (a_2, p_2), returning *(price, positive)* and *(service, negative)*.

During the Aspect Term Sentiment Classification task, we observe that current models tend to overemphasize the mere presence of the aspect term and the majority bias related to the term, neglecting the broader context of the sentence when determining sentiment. This tendency may not be readily apparent when utilizing SemEval datasets, as these are hand-crafted where training and test sets exhibit similar distributions. However, when deploying real-world datasets exhibiting highly skewed

or missing labels for specific aspects, the vulnerability of current ABSA models becomes more evident. These models tend to memorize the associated word and repeatedly predict the sentiment label previously associated with it. To mitigate this issue, we introduce the MaskedABSA method. This technique involves replacing aspect terms within a sentence with a placeholder token, "[MASK]", thereby redirecting the model's focus towards the context of the sentence.

The procedure modifies the original sentence from $S_i = \{w_1, w_2, w_3, \ldots, w_n\}$ to $S_i' = \{w_1, a_1, w_3, \ldots, a_k, \ldots, w_n\}$ considering both the input S_i and aspect terms A_i. For the Mask method, we construct masked sentences such as $M_{i_1} = \{w_1, \text{[MASK]}, w_3, \ldots, w_n\}$ and $M_{i_k} = \{w_1, w_2, w_3, \ldots, \text{[MASK]}, \ldots, w_n\}$ from S_i', where each aspect term a_k is replaced by "[MASK]" and for each aspect a new sample M_{i_k} is created. The model, denoted as F_{MASC}, is then tasked with predicting the polarity of the term represented by "[MASK]" in each sentence. The model returns a polarity value p_k. Thus, we can express the function of the model as $F_{MASC}(M_{i_k}) = p_k$.

5.1.2 Weakly Labeled Dataset

One effective strategy to address the scarcity of ABSA datasets and the challenges of bias involves using weak supervision combined with social network structures, such as retweet networks [113]. Analyzing the patterns of retweets can infer the sentiment and aspect orientations prevalent within specific user groups or topics without requiring detailed sentence-by-sentence manual annotation. This approach allows for creating large-scale, weakly labeled datasets that can be refined using machine learning models and further validated through targeted reviews.

A well-defined codebook, offering guidelines for detecting and annotating stances on key aspects according to their distinct community orientations, significantly enhances the efficiency and consistency of the weakly labeled data production process.

Several researchers have utilized the Louvain community detection algorithm on social media networks [114],[115],[116], to extract non-overlapping communities. It was also observed that [117] users within the same Louvain community tend to share not only common interests but similar viewpoints and stances on popular issues and

entities. Our approach harnesses per community homogeneous viewpoints, captured in a codebook, to generate vast amounts of weakly labeled datasets for ABSA tasks.

Incorporating weakly labeled datasets into ABSA pipelines also facilitates scalable model training, where transfer learning techniques can be employed to fine-tune large-scale pre-trained models on domain-specific data. This combination of weak supervision and transfer learning ensures that the inherent label noise is counterbalanced by advanced models' robust feature extraction capabilities, leading to improved sentiment and aspect detection accuracy. Furthermore, this process reduces the dependence on scarce high-quality labeled datasets, enabling researchers to focus their annotation efforts on resolving critical ambiguities in the data.

5.1.2.1 U.S. Race Relations Dataset

Our dataset, derived from a real-world case, contains a collection of 9,084,824 tweets from January 1, 2022 to May 31, 2023, related to race relations in the U.S. The dataset involves 296,540 users and includes a retweet network with 738,012 edges, depicted in Figure 6.1. To collect the dataset, we used keywords related to U.S. race relations, specifically focusing on five topics - Black Lives Matter (BLM), All Lives Matter (ALM), Issues, Races, and Religion. Weakly labeled data is stance annotated data, leveraging shared community biases as recorded on the codebook.

The primary advantages of producing weakly labeled datasets using our approach are its cost-efficiency and scalability, making it an appealing option in scenarios where acquiring comprehensive, fully labeled datasets is economically or logistically not feasible.

5.1.2.2 Camp Labeling of Users

The retweet network graph shows a polarized network comprising two camps with seven sub-communities. Each sub-community is depicted in a different color in Figure 6.1.

An experimentally determined edge weight threshold is applied to partition the net-

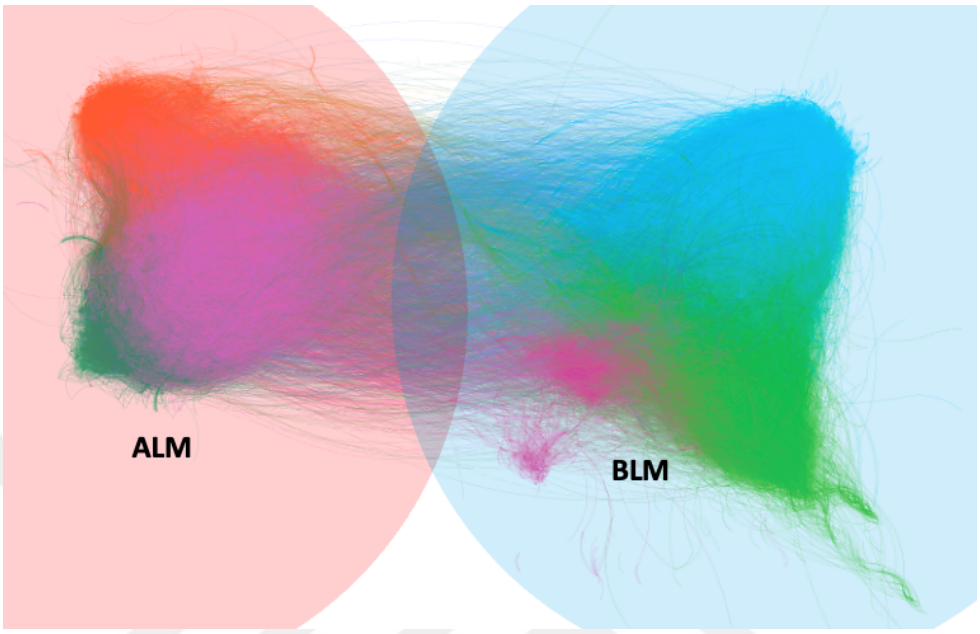


Figure 5.2: Visualization of polarized camps in U.S. Race Relations

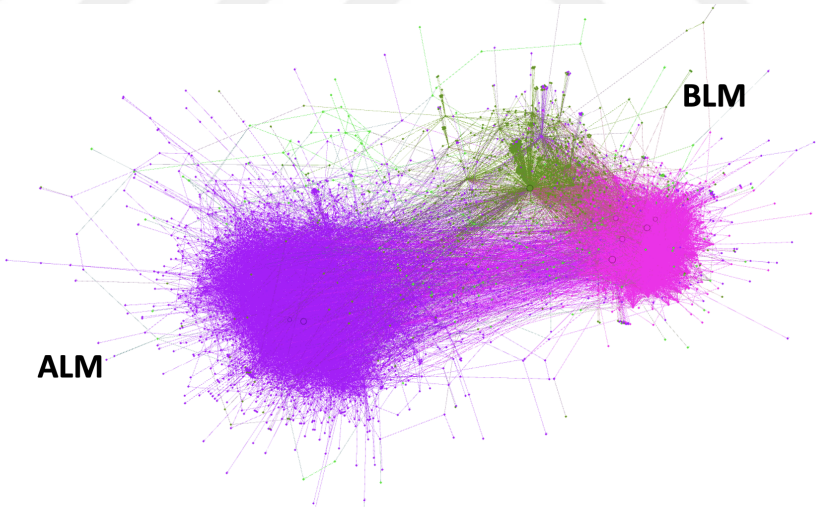


Figure 5.3: Only Barrier-crossing users' network from U.S. Race Relations dataset

work into a pair of polarized camps. A panel of experts labels these camps as either pro-BLM or pro-ALM based on an analysis of their most viral 1,000 tweets, examined community by community. In social media analytics, users' communication patterns can be categorized as either barrier-bound or barrier-crossing. Barrier-bound partisan users interact only within their camps, whereas barrier-crossing users attract positive engagement from both camps. The barrier-crossing users' network is shown in Figure 5.3. Previous studies [118],[119] have shown that content amplified by barrier-bound users reinforces existing biases within their groups, creating echo chambers that further polarize opinions. Our approach leverages the consensus among barrier-bound users who exhibit explicit partisan biases to develop a codebook and rapidly generate vast amounts of weakly labeled training datasets for MaskedABSA models.

5.1.2.3 Bias and Sentiment Coding

We extracted all noun phrases and named entities from our corpus using weak labeling for significant aspect terms. Our codebook encompasses the most frequent and discriminating unigrams, bi-grams, and other n-grams, including all unambiguous bigrams such as "mental health" that could be overshadowed by more generic high-frequency unigrams like "health."

Subsequently, three domain experts independently coded the noun phrases and named entities for their shared stances within the BLM and ALM camps. These experts then convened in a panel to discuss and reconcile all their differences, aiming to achieve a consensus on the codebook entries.

From the top 3,000 phrases identified, 1,789 are determined to be non-ambiguous. The sentiment column reveals 496 phrases with negative connotations and 257 with positive ones. In terms of bias, 234 phrases are coded as "anti" and 160 as "pro" within the BLM camp; for the ALM camp, 177 phrases are "anti" and 90 are "pro".

The resulting dataset comprises weakly labeled tweets aimed at dissecting sentiment and ideological biases linked to specific camps and their sub-communities. Annotations comprise both aspect and stance as depicted in Figure 5.4.



Figure 5.4: An example of a stance annotated sentence from the BLM camp.

5.1.2.4 Preparing U.S. Race Relations and Politics Datasets

From 62,196 samples, we separated two distinct sets of data focused on a pair of different themes: one on race-related issues and the other on politics-related issues and entities. Initially, we identified 1,206 terms associated with race-related topics and 593 terms with politics-related topics. We then created two separate collections of tweets: one matching terms related to race relations and another matching terms related to U.S. politics. We ensured these datasets did not overlap; any tweet that included terms from both categories was left out. After organizing these groups, we balanced the training and testing datasets for race relations and politics themes. We focused on ensuring fair representation from the All Lives Matter (ALM) and Black Lives Matter (BLM) camps. This approach led to the identification of 6,760 tweets for the politics dataset and 84,518 tweets for the race relations dataset.

5.2 Experiments and Evaluations

The experiments in this section tests our MaskedABSA approach using the SemEval Dataset and real-world datasets from Twitter.

5.2.1 SemEval Datasets

For the SemEval datasets, we used four training datasets: Laptop14, Restaurant14, Restaurant15, and Restaurant16. These datasets each include training and test subsets, allowing us to conduct our experimental evaluations comprehensively.

The data illustrated in Table 5.1 and Tabel 5.2 demonstrate that the MaskedABSA model surpasses alternative models in terms of performance across all datasets, showing exceptional robustness even in cross-domain testing scenarios.

Table 5.1: ATSC accuracy results (%) on SemEval datasets compared with Masked-ABSA models.

Model	Laptop14	Res.14	Res.15	Res.16
BART-ABSA-ALSC[50]	76.76	*87.29*	76.49	89.10
Dual-MRC[48]	75.97	82.04	75.08	-
InstructABSA2 [52]	*81.56*	85.17	*84.50*	*89.43*
MaskedABSA	**86.24**	**87.65**	**91.53**	**94.83**

Table 5.2: Cross-domain ATSC results (%) on SemEval datasets comparing with the top model in the literature.

Train	Test	Model	Accuracy
Restaurant 14	Laptop14	InstructABSA	82.44
		MaskedABSA	**84.79**
Laptop 14	Restaurant 14	InstructABSA	80.56
		MaskedABSA	**85.32**
Restaurant 15	Hotel 15	InstructABSA	89.74
		MaskedABSA	**91.37**

5.2.2 U.S. Race Relations and Politics Datasets

To ensure our dataset is similar in size and scope to the SemEval datasets, we randomly chose 3,000 tweets for training and 1,000 tweets for testing. While selecting these tweets for training and testing, we focused on balancing the ALM and BLM camps. Due to the nature of Twitter, where negative messages often spread more effectively [120],[121], our dataset ended up having more negative content than positive. The details and key metrics of our dataset are shown in Table 5.3.

Table 5.3: Race and Politics datasets distribution.

Races Dataset	Positive	Negative	Total	Politics Dataset	Positive	Negative	Total
Train	885	2115	3000	Train	873	2122	2995
Test	309	699	1000	Test	257	751	1008

We started by developing two models, one trained on the U.S. race relations dataset and the other on the U.S. politics dataset. The performance of each model was measured, with the race relations model achieving an F1 score of 71.92 and the politics model scoring 72.10, as shown in Table 5.4.

For cross-domain validation tests, we tested the performance of each model using the dataset from the other domain. The results showed only small drops in their corresponding F1 scores, only 2.68 points for the U.S. politics dataset and 2.70 points for the U.S. race relations dataset. These experiments indicate that the MaskedABSA model maintains strong cross-domain performance, as shown in Table 5.4.

Table 5.4: Cross-domain test results on the U.S. Race Relations and Politics datasets using MaskedABSA.

	Single		Cross		Differences
	Test	F1	Test	F1	
Race	Race	71.92	Politics	69.24	-2.68
Politics	Politics	72.10	Race	69.40	-2.70

5.2.3 Bias Test

Aspect-terms-bias is a phenomenon where a model favors specific stances for specific terms. To tackle this problem, we developed the masking method, where context is forced to determine the stance instead of the aspect itself. To evaluate the efficacy of our approach, we conducted a bias test analyzing the accuracy of stance detection corresponding to messages carrying minority category biases. For instance, if the word "gun control" appears ten times in the dataset with eight occurrences labeled as anti/negative and two as pro/positive, we aimed to determine the accuracy of ABSA on the minority stance category tweets. In this test, we evaluated the performance of the InstructABSA model in comparison to the MaskedABSA model. The results demonstrated that our MaskedABSA model achieved an average accuracy of 45.6% and 57.7% in the politics and race-related models, significantly surpassing the InstructABSA's accuracies of 2% and 8% only, respectively. This experiment reveals a

fundamental limitation of state-of-the-art ABSA models: a susceptibility to memorizing majority category stance labels for aspects, even in contexts that unambiguously signal a minority category stance. This highlights a critical weakness in their ability to accurately discern and accurately represent the perspectives of less prevalent groups.

5.2.4 New Terms Test

To assess the stance detection accuracies for the previously unseen, new aspect terms issue, we designed an experiment to gauge the effectiveness of the MaskedABSA. We analyzed the model's performance with new aspect terms across U.S. race relations and politics datasets. Our approach involves replacing the queried aspect with a new term that does not appear in the training set. As a result, the aspect terms used in the test datasets were completely new to the models, having not been introduced during their training. This testing procedure was replicated for both the race relations and politics datasets to measure the performance of the InstructABSA model. Despite the experimental conditions, the MaskedABSA model's accuracy remained the same, as it conceals the queried aspect term during prediction, rendering the novelty of the term inconsequential. Additionally, we applied this methodology to the SemEval dataset, introducing new terms to challenge the model further. The results, documented in Table 5.5, clearly illustrate the MaskedABSA model's enhanced ability to handle new terms compared to the other models.

Table 5.5: Accuracy with replacements of new terms.

	Laptop 14	Res. 14	Res. 15	Res. 16	Race	Politics
Instruct (%)	50.00	71.23	55.08	64.40	53.00	61.84
Masked (%)	**86.24**	**87.65**	**91.53**	**94.83**	**71.92**	**72.10**

5.3 Discussion

This chapter focuses on exploring bias in sentiment analysis and the challenges posed by echo chambers in social media platforms. By addressing these issues, this work demonstrates the potential of advanced aspect-based sentiment analysis (ABSA) techniques to deliver fairer, more accurate, and scalable insights. The proposed Masked Aspect Sentiment Classification (MASC) framework introduces significant advancements in mitigating bias while maintaining the granularity required for effective sentiment evaluation.

One of the chapter's key contributions is the integration of masking techniques to handle bias in aspect-based sentiment analysis. Biases, whether introduced by training data or inherent in language models, often skew sentiment predictions, particularly in highly polarized contexts like politics or social issues. The proposed framework systematically masks potentially bias-inducing terms, allowing the model to focus on more neutral and relevant features. The results indicate substantial fairness and accuracy improvements across benchmark datasets and real-world examples, validating the framework's efficacy.

The use of weakly labeled datasets further highlights the scalability of the MASC framework. By leveraging heuristics and semi-supervised techniques, this study demonstrates that large-scale datasets can be effectively utilized without extensive manual annotation. This approach addresses the challenge of data scarcity and ensures that the framework remains adaptable to evolving topics and sentiments.

A notable finding of this chapter is the framework's ability to detect and analyze sentiment in echo chamber environments. The experimental results reveal how biases manifest differently across polarized groups, emphasizing the importance of addressing these biases to enable meaningful cross-group comparisons. Moreover, the framework's performance on unseen terms and new topics demonstrates its robustness and potential for generalization in dynamic social media landscapes.

Despite these achievements, some limitations warrant discussion. First, while the masking strategy effectively reduces bias, it may inadvertently omit terms that carry critical contextual information. Future work could explore adaptive masking ap-

proaches that dynamically adjust the masking level based on the context of the analyzed data. Additionally, incorporating attention mechanisms to weigh the importance of masked terms could enhance the interpretability and precision of the model.

Second, while scalable, the weakly labeled datasets may introduce noise and uncertainty into the training process. Although the results indicate strong overall performance, further refinement in the labeling heuristics could enhance data quality and improve model outcomes. Developing methods to quantify and mitigate noise in weakly labeled datasets represents a promising avenue for future research.

Another challenge lies in extending the framework to multilingual and multimodal contexts. While the current implementation focuses on textual data in English, the growing diversity of social media content necessitates support for multiple languages and media types, including images and videos. Addressing these dimensions would expand the framework's applicability and relevance to global social media ecosystems.

In conclusion, this chapter provides a robust framework for mitigating bias and enhancing sentiment analysis in polarized and large-scale contexts. The study sets a new benchmark for fairness and accuracy in ABSA by introducing innovative masking techniques and leveraging scalable data labeling methods. Future research could focus on extending these methods to broader contexts, incorporating multimodal data, and refining adaptive strategies for dynamic and real-time sentiment analysis. The insights presented here contribute to developing more equitable and effective tools for understanding social media dynamics.

CHAPTER 6

TOWARDS A PROGRAMMABLE HUMANIZING AI THROUGH SCALABLE STANCE-DIRECTED ARCHITECTURE

The increasing use of generative artificial intelligence (AI) in digital communication highlights the need for AI systems prioritizing safety and ethics. This chapter introduces a novel pipeline to curb the generation of toxic narratives by training more humanized Large Language Models (LLM). Community resilience building literature [122], indicates that social cohesion, where communities foster a sense of belonging and shared values, acts as a bulwark against toxic speech. Unlike toxic, polarizing speech, cohesive language can empower individuals to understand the harm of hateful content and stand together against it. Therefore, we develop an approach to fine-tuning generative AI on positive, core human values and common good, reinforcing arguments to promote healthier, more humanizing, and constructive discourse [123].

Existing Reinforcement Learning from Human Feedback (RLHF) [124] based approaches are criticized for being expensive, time-consuming, and prone to annotator bias, often leading to models that shy away from generating content on sensitive topics. This type of training leads to either overarching guardrails, which results in refusals to generate content, or overly contentious rhetoric whenever the prompt can escape the guardrails of the LLM, particularly when dealing with sensitive or controversial issues.

Our study proposes a scalable processing pipeline that balances cost efficiency and comprehensiveness. This is accomplished by the creation of a *'concise'* codebook that includes the most informative terms, which are systematically coded for sentiment (pos/neg), stance (pro/anti), and societal considerations (common good vs. othering) by a select panel of experts. This methodological approach enables the generation

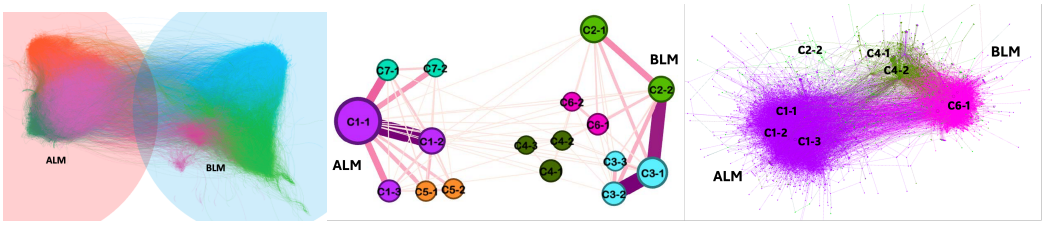


Figure 6.1: Retweet Network Graph of Seven Major Communities, 17 Sub Communities, and Barrier-crossing Users in U.S. Race Relations network labeled with corresponding sub-community IDs.

of extensive amounts of stance-aware training data towards programmable, stance-directed civilizing and humanizing AI.

ABSA plays a vital role in understanding the fine-grained sentiments expressed by users. Although stance detection is defined differently in different application settings, the most common definition is "automatic classification of the stance of the generator of a piece of text, towards a target, into one of these three classes: Favor, Oppose, Neither."

To evaluate the efficacy of the above concepts, we utilize a real-world Twitter dataset comprising 15,914,812 tweets from 296,540 users and 738,012 shares covering roughly one month between 4/18/2023 and 5/31/2023 on the U.S. Race Relations-related topics.

Retweeting (or sharing) is a widely adopted feature on Twitter for spreading and making information go viral [125]. A recent study on political discourse on social media [110] confirms that Twitter users are exposed mainly to opinions that agree with their own, and partisan users enjoy higher centrality and content endorsement. In other words, camps and echo chambers are very real on Twitter.

Initially, we utilized community detection algorithms as in [115] on the retweet network to expose a pair of polarized camps alongside their community/sub-community network structures (see Figure 6.1). In order to discriminate between partisan users and their more moderate, as discussed in previous chapter, we identified two types of users in both camps: partisan *'barrier-bounds'* vs. more moderate *'barrier-crossers'*. Barrier-crossing users are defined as those in each camp that interact (i.e., by retweeting or being retweeted-by) with users from their own and with users from the other

camp. Barrier-bound users are defined as those who interact with users from their own camp only [126].

In the following sections, we outline the architecture where our experiments demonstrate the success of our proposed architecture in curbing toxic content generation. The resultant model significantly reduces the propensity of an unrestricted LLM to produce toxic content by 85%. The experimental dataset and models used in our evaluations are shared online on GitHub tweetpie/stance-directed-humanizing-ai.

6.1 Dataset and Camp Labeling

This section describes the datasets and the methodology employed to label users into distinct camps. Our labeling process combines community detection algorithms with expert panel reviews to ensure accurate identification of polarized camps within the dataset.

6.1.1 Camp Labeling of Users

We employed the Louvain community detection algorithm [127], a widely used method for uncovering modular structures in large networks. This algorithm identified seven major communities with 17 sub-communities, represented as nodes of varying sizes and depicted in distinct colors in Figure 6.1. The Louvain method's modularity optimization ensures that densely connected nodes are grouped, revealing underlying social structures within the network.

To determine polarized camps, we introduced an experimentally defined edge weight threshold, which partitioned the network into two overarching groups. Our expert panel carefully reviewed the top 1,000 most viral tweets from each sub-community to assign labels. These viral tweets, identified based on engagement metrics such as retweets and likes, were analyzed for sentiment, hashtags, and language patterns. Through this inspection process, the panel labeled the two polarized groups as pro-Black Lives Matter (BLM) and pro-All Lives Matter (ALM).

The approach provides a robust method for classifying users, balancing the objectiv-

ity of network-based algorithms with the contextual understanding of human experts. The resulting camp assignments were further validated by comparing the language and hashtag usage trends across the two camps, ensuring that the communities' ideological differences were reflected in their online activity.

6.1.2 Bias, Sentiment and Common Good Coding

We extracted all noun phrases and named entities from the corpus using Named Entity Recognition (NER) and Noun Chunk processing. Our codebook includes unigrams, bigrams, and other n-grams. We noted that unigrams such as "health" might overshadow more informative bigrams like "mental health," even if "mental health" appears more frequently. To address this, we implemented a discounting mechanism, adjusting unigram frequencies by subtracting the frequencies of overlapping longer n-grams, provided their cumulative frequency exceeds that of the unigram. Finally, we used the "elbow method" to determine the optimal cutoff frequency for identifying the most informative phrases for the codebook.

Next, three domain experts were asked to code the noun phrases and named entities independently for *ambiguity*, *pos/neg sentiment*, *pro/anti bias* towards BLM or ALM camps, and societal considerations associated with *common good vs. othering*. Following that, three experts were put into a panel to discuss and reach a common consensus by resolving all of their disagreements and finalizing the design of the codebook.

The top 3,000 phrases identified through the analysis were subject to coding, with 1,789 deemed as non-ambiguous. The sentiment column has 496 phrases with negative connotations and 257 with positive ones. When examining bias, 234 phrases were coded as "opposed", with 160 as "supported" by the BLM camp, whereas 177 phrases were coded as "opposed", with 90 as "supported" by the ALM camp. Also, 214 phrases were coded as "common good" versus 567 phrases as "othering/divisive". Furthermore, 85 phrases were coded as core human values categorized under the following eight topics: Family, Life, Truth, Nation, God, Democracy, and Justice. Figure 6.2 shows the phrase clouds retrieved with the assistance of the codebook.



Figure 6.2: Word cloud of the codebook displaying (left-to-right) positive/negative phrases, common good/public interest and othering/divisive phrases. Sizes represent their frequencies.

The codebook and its analysis offer a powerful lens for understanding the complexities of U.S. race relations debates on social media. By delving beyond surface-level sentiment and identifying potential biases and divisive terms, our approach aims to enable the synthesis of more reliably labeled training datasets with less effort to develop a programmable Humanizing AI that can prioritize core human values and common good related narratives.

6.1.3 Weakly Labeled Dataset

Weakly labeled data has not been explicitly tagged by manual effort. However, this type of data cannot be considered entirely unlabeled either since it is developed under certain verifiable assumptions across the dataset. Weakly labeled datasets are preferred due to their cost-effectiveness and scalability, particularly in applications where fully labeled data is prohibitively expensive or logistically impossible. These datasets leverage weaker forms of supervision—such as camps, keywords, or even noisy labels—to train AI models, which brings the ability to rapidly amass large volumes of training data, potentially improving the generalization capabilities. Figure 6.3 shows the curation process for the weakly labeled dataset.

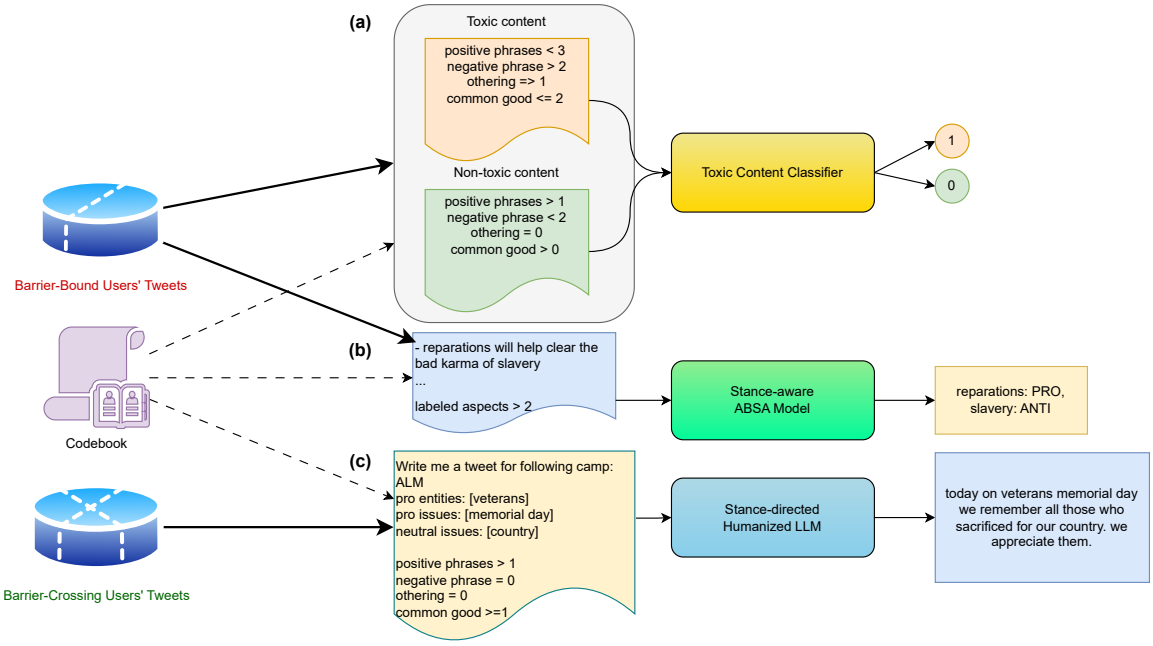


Figure 6.3: Weakly labeled dataset preparation rules based on the codebook for the language models with running examples.

Barrier-bound users, characterized by their tendency to produce more partisan content, express a higher prevalence of "othering" language in their messages than barrier-crossing users. This observation underscores the importance of distinguishing between these two user groups when curating datasets for AI training. Our approach leverages barrier-bound users' tweets to train our ABSA-based stance-aware reader and toxicity classifier models. The tweets from barrier-crossing users are reserved for testing the efficacy of these models. In earlier research [118], it is observed that amplification (i.e., by shares, quotes, or likes) of partisan barrier-bound messaging within their own camps uses social media echo chambers to polarize the network effectively. At the same time, amplification of barrier-crossing messaging within their own camps engenders moderating effects.

To develop an effective toxic content classifier, we constructed a rule-based dataset for identifying and labeling tweets as toxic or non-toxic. Non-toxic content was identified based on the presence of at least two positive phrases, at most one negative phrase, the absence of othering/divisive language, and the inclusion of at least one common good/public interest phrase. Conversely, toxic content was defined by the presence of at most two positive phrases, at least three negative phrases, at least one othering/divisive phrase, and at most two common good/public interest phrases. The resulting dataset comprised 33,787 training tweets, with 12,757 tweets originating from the ALM camp and 21,030 tweets from the BLM camp. Within the training set, 29,005 tweets were labeled toxic, while 4,782 were labeled non-toxic. The proportion of non-toxic content was slightly higher in BLM tweets, accounting for 17.4%, compared to 13.2% in ALM tweets.

We prepared another weakly labeled dataset for the ABSA training task to capture various aspects of sentiment and camp bias within tweets. The dataset combined sentiment and camp-bias labels, prioritizing camp bias when present and defaulting to sentiment labels in cases where camp bias was neutral. After labeling the data, we filtered the tweets to include only those containing at least three labeled phrases to ensure robust training samples. The dataset comprises 62,196 training samples. An illustrative example from the BLM camp for the sentence "*reparations will help clear the bad karma of slavery*" is represented as "*[reparations/PRO] will [help/PRO] clear the bad karma of [slavery/ANTI].*"

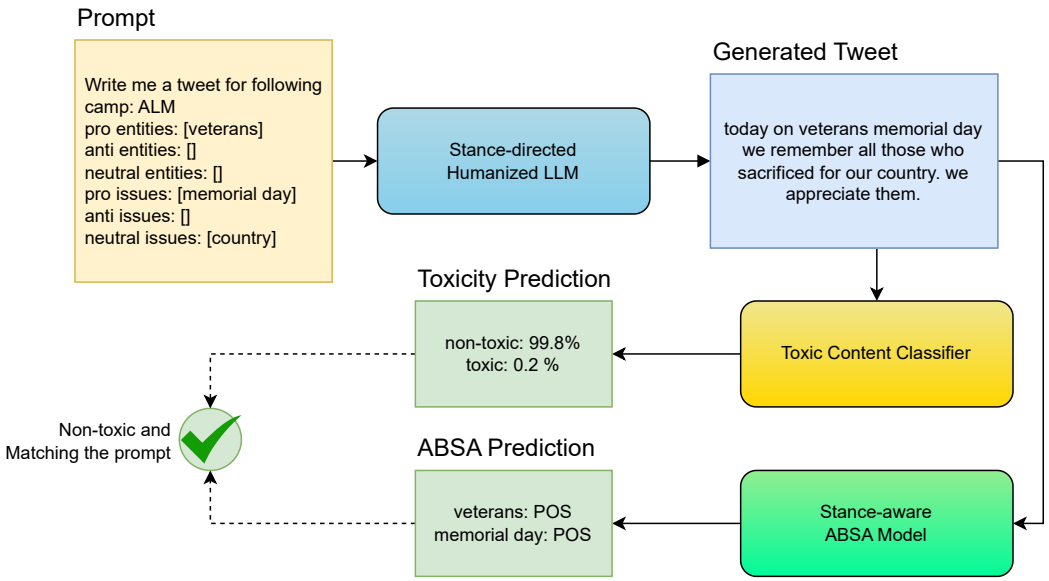


Figure 6.4: The process of generating a non-toxic message thru a stance-directed prompt on certain entities and issues.

Following that, we adopted barrier-crossing users' content to develop a programmable stance-directed generative AI by employing two different strategies. The first strategy involved constructing a dataset devoid of explicit toxic content to guide the generation of a socially responsible and inclusive model. Specifically, tweets from barrier-crossing users were selected where there was no mention of any othering/divisive category or any negative sentiment category phrase, and there was the presence of at least one common good/public interest category phrase. The second strategy utilized an unfiltered barrier-crossing subset of the dataset, allowing for greater expressivity in content without imposing any explicit rules, thus exposing the model to a broader discourse.

Ultimately, 7,966 tweets matching the first strategy criteria were identified, with 57.3% originating from the BLM camp, underscoring the diversity of perspectives captured within the dataset. For the second strategy, a combination of half of the tweets from the first strategy's dataset and a similar amount of random tweets from the barrier-crossing users were utilized to ensure a similar sample size for fine-tuning in our experiments.

6.2 System Architecture

Our system architecture, depicted in Figure 6.4, demonstrates generating non-toxic messages. It highlights stance-directed prompts on selected entities and issues to guide the message creation. However, the inherent unpredictability and potential for generating hallucinations with LLMs necessitate a post-filtering process to ensure the generated contents' appropriateness.

Following the content generation phase, *toxic content classifier* and *ABSA* models were employed to inspect the generated tweets' toxicity and stances as a post-filter. Implementing these models as a final filter ensured that only non-toxic categorized tweets matching prompted stances could be released to their intended audiences.

6.2.1 Stance-Directed Humanized LLM

We utilized Flan-T5 [128] to fine-tune humanized and unrestricted LLMs in our application. For the humanized LLM, we fine-tuned Flan-T5 using a dataset of barrier-crossing user tweets that adhered to specific guidelines outlined above as Strategy 1 and 2, ensuring fairness and inclusivity in the generated content. The model was trained to generate tweets framed around specific camp biases while avoiding divisive language and maintaining a balanced representation of diverse viewpoints. Conversely, the unrestricted LLM was fine-tuned using a combination of tweets from the humanized dataset and a random subset of tweets from the barrier-crossing users, allowing for a broader exploration of language patterns and expressions.

The model is prompted through a programmable template requesting: "*Write me a tweet for following camp: '<ALM or BLM>', framing pro entities: [], anti entities: [], neutral entities: [] based on following issues pro issues: [], anti issues: [], neutral issues: []*." The distinction between entities and issues lies in the type of noun phrases mentioned in the codebook, with entities being standardized named entities and issues being topic-coded noun phrases. Including neutral aspects aims to help the model mention specific topics without necessarily taking a clear stance as pro or anti.

6.2.2 Aspect-Based Sentiment Analysis

InstructABSA [52] builds upon the Tk-Instruct model [129], a framework designed for instruction tuning LLMs to perform better on domain-specific tasks. InstructABSA extends this concept specifically to the ABSA domain by leveraging the inherent capabilities of Tk-Instruct to grasp sentiment expressions related to different aspects of a subject. The core mechanism of InstructABSA involves instruction tuning, where the model is fine-tuned with specific instruction prompts related to ABSA tasks. These prompts encompass task definitions and examples representing positive, negative, and neutral sentiments. This method aims to refine the model's ability to identify, extract, and classify sentiments associated with particular aspects of a text.

InstructABSA's prompts are structured to include a clear task definition followed by exemplars showcasing how to analyze sentiments tied to different aspects. These prompts are tailor-made for each ABSA subtask, ensuring that the model receives direct and relevant guidance on handling the intricacies of sentiment analysis at the aspect level. The output of the ABSA task is the list of extracted aspects alongside corresponding stances. An example stance-directed prompt and its generated output can be seen in Figure 6.4.

6.2.3 Toxic Content Detection

Our pipeline leverages the BERT [43] classifier in toxicity/non-toxicity detection, which has the advantage of a deep understanding of language context. To enhance BERT's performance in toxicity detection, we fine-tuned a pre-trained hate speech detection model [130]. This model comprises numerous examples of hate speech on race, religion, gender, sports, and politics, enabling the model to learn the specific characteristics and linguistic patterns associated with offensive content. The fine-tuning approach significantly improved the model's ability to detect a wide range of toxic language, from overt hate speech to more subtle forms of toxicity.

The input to the BERT-based toxicity detection model consists of raw text data, typically sentences or paragraphs extracted from social media posts, comments, or other user-generated content. The model's output is a classification score that indicates the

likelihood of the input text being toxic. This score is derived from the final layer of the BERT model, which is a dense layer customized for the binary classification task of identifying toxic versus non-toxic content.

6.3 Experimental Evaluation

Our experimental evaluations are predicated on a novel approach to dataset segmentation, wherein tweets from barrier-bound users are allocated for training, and tweets from barrier-crossing users are reserved for testing. This division is designed to simulate a realistic scenario where the model is trained on a homogeneous group of users but tested on diverse users that span different communities, thereby assessing the model's capabilities across various social orientations.

To mitigate the risk of the toxic content classifier model overfitting and memorizing linear relationships due to specific keywords, we employed a codebook-based *phrase removal* within the tweets. By removing phrases with a certain likelihood, we aimed to reduce the undue influence of high-frequency words found in the codebook on the classification outcomes. This enables our model to tackle the problem of high-frequency keywords such as "family" skewing the detection of toxicity, even in fundamentally toxic content. Removing high-frequency phrases entirely would wipe out the valuable insights that could be derived from our codebook, whereas retaining them would lead to overfitting through memorization. After experimenting with removal probabilities of 0, 30, 60, 80, 98, and 100 percent, we found that a removal probability of 80 percent offered the most accurate results. To enhance the diversity of our dataset and counteract potential biases, we integrated it with the hate speech dataset from the hate speech detection study [130], which comprises three categories: neutral, offensive, and hate. We merged the offensive and hate categories under a single "toxic" label while designating the neutral category as "non-toxic."

A gold test dataset of 1,000 expert-labeled tweets was also developed to evaluate the performance of the toxic content classifier. For this evaluation, toxic content was defined as stereotyping, fostering division, employing offensive language, or demeaning/dehumanizing individuals or groups. We employed Krippendorff's alpha

Table 6.1: Performance of the classifier on toxicity detection

Test Dataset	Precision	Recall	F1
Gold Dataset	0.81	0.81	0.81
LLM Generation	0.77	0.76	0.76

Table 6.2: Comparable reduction in toxicity on sensitive domain

Model	Sensitive	Toxicity Reduction
Unrestricted	73%	-
Humanized	27%	64%
Humanized + Filter	11%	85%

coefficient to measure inter-annotator agreement, achieving a strong nominal score of 0.73, demonstrating high consistency in the assigned labels.

The final phase of our experimental setup involved applying the toxic content classifier to the LLM-generated tweets. To ensure the integrity of our data, an expert panel was engaged to label the generated tweets at an inter-rater reliability level of a nominal score of 0.75 using Krippendorff's alpha. By comparing the classifier's predictions against the expert-coded labels, we aimed to quantify our pipeline's effectiveness in a real-world scenario.

Our toxic content classifier model achieves an accuracy between 76% and 81% in accurately categorizing tweets as toxic or non-toxic, as detailed in Table 6.1. Our assessment revealed that the ABSA model demonstrates proficiency in extracting 63% of the issues and entities from the generated text. Moreover, the model accurately matches 83% of the identified stances with the directives given in the prompt, indicating a high level of alignment.

The unrestricted LLM generates toxic content 73% of the time. Significantly, our humanized LLM achieved a 64% reduction in toxicity when responding to prompts on sensitive issues and an impressive 83% on general domain-related discourse. Even more compelling, integrating the toxic content classifier with the humanized LLM

further boosted these reductions to 85% for the sensitive issue-related generation and 86% for the general domain-related generation, as shown in Table 6.2.

6.3.1 Comparison with pre-trained LLMs

The findings presented in this section illustrate the responses of popular LLMs to generating tweets on sensitive topics with left or right ideological leanings based on 30 balanced prompts. This analysis highlights the capabilities and limitations of contemporary LLMs in content moderation and bias mitigation. The humanized LLM model achieved a 97% response rate, with only one instance of toxicity, showcasing its robustness in handling sensitive content with minimal adverse effects. In comparison, ChatGPT-4o produced relevant responses 60% of the time without any toxic content, indicating its tendency to avoid sensitive entities like racial groups when prompts request anti-mentions. Gemini and Claude showed significant reluctance to engage with right-wing ideologies, with Gemini refusing 63% of the time and Claude 40%. However, both models were more responsive to left-leaning prompts, indicating a variance in engagement based on ideological context. Additionally, replacing right-wing entities and ideologies with their left-wing counterparts resulted in increased responsiveness, underscoring the impact of annotator bias.

Mistral showed a higher related response rate of 73%, comparable to the humanized LLM model, but produced abstract and less domain-specific tweets. Like the humanized LLM model, this model had a single instance of toxic content generation, underscoring the challenge of maintaining non-toxic outputs across diverse prompts. Despite these differences, all LLMs managed to generate predominantly non-toxic tweets. The comparison underscores the humanized LLM model's efficacy in handling sensitive topics with high relevance and low toxicity, outperforming other LLMs in consistency and engagement. Detailed prompts and outputs are accessible in the public GitHub repository.

6.4 Discussion

This chapter introduces a novel stance-directed AI architecture designed to humanize content generation by aligning outputs with user preferences and perspectives. By addressing the limitations of generic AI models in producing context-aware, stance-driven, and human-like content, the proposed framework marks a significant step forward in advancing personalization and engagement in artificial intelligence applications.

A central contribution of this chapter is the integration of stance detection with content generation, enabling AI systems to align their outputs with user-defined stances while maintaining coherence and relevance. The experimental results demonstrate that the proposed architecture effectively generates personalized content that reflects user preferences and ideological alignments. This capability is particularly impactful in contexts requiring sensitive or nuanced communication, such as political discourse, marketing, or customer engagement.

A scalable, programmable architecture ensures adaptability across various domains and user requirements. The system's modular design allows for seamless incorporation of additional components, such as sentiment analysis or toxicity detection, further enhancing its utility. Moreover, RLHF underscores the importance of iterative improvement in aligning AI outputs with human expectations, fostering a more interactive and responsive content generation process.

Another notable achievement is the framework's ability to mitigate the risk of toxic or biased content generation. By incorporating toxicity detection mechanisms and leveraging weakly labeled datasets, the model balances stance alignment with ethical considerations, ensuring that generated content adheres to principles of fairness and inclusivity. This dual focus on personalization and ethical content generation addresses a critical challenge in deploying AI systems in real-world scenarios.

However, several challenges and limitations remain. One of the primary limitations is the dependency on high-quality labeled data for stance and sentiment classification. While weakly labeled datasets reduce reliance on manual annotation, the potential for noise and inaccuracies in the training data may impact model performance. Fu-

ture research could explore advanced data augmentation and self-supervised learning techniques to enhance data quality and diversity.

Another limitation is the model's sensitivity to the granularity of stance definitions. The framework's effectiveness depends on the precision of stance labels and the clarity of their definitions. Ambiguity or inconsistency in stance definitions may lead to misaligned outputs, particularly in highly polarized or subjective contexts. Developing dynamic stance refinement methods and incorporating user feedback could address this issue.

Additionally, the computational complexity of scaling the architecture to real-time applications poses a challenge. While the framework demonstrates strong performance in controlled experimental settings, optimizing its efficiency for deployment in resource-constrained environments remains an area for future exploration. Techniques such as model compression and distributed processing could help mitigate these challenges.

In conclusion, the stance-directed architecture presented in this chapter represents a significant advancement in programmable, humanized AI systems. By bridging user preferences, ethical considerations, and scalable design, the framework offers a robust solution for personalized content generation in diverse applications. Future research directions include refining stance classification, expanding multimodal capabilities, and optimizing real-time performance. The contributions of this chapter lay the groundwork for further innovation in creating AI systems that are not only effective but also ethical and responsive to user needs.

CHAPTER 7

CONCLUSION

This thesis has presented a comprehensive exploration of the integration of AI with personalization techniques to harness social media's vast potential for targeted marketing. The study has addressed critical challenges in user classification, personalized content generation, and sentiment analysis, contributing to the broader field of AI-driven human-centric solutions. Below, we summarize this research's key findings, contributions, and implications alongside opportunities for future work.

The primary contributions of this thesis are as follows:

1. **User Classification:** This work introduced a robust approach for classifying Twitter accounts as individual or organizational based solely on account metadata. The proposed method simplifies preprocessing by minimizing reliance on tweet data and offers a language-independent solution.

2. **Personalized Marketing:** A framework was developed for generating coherent and personalized landing pages informed by user-generated content and preferences. The proposed method integrates topic modeling and text generation to create relevant and engaging marketing materials.

3. **ABSA:** This research advanced sentiment analysis techniques to address challenges such as bias and echo chamber effects in social media data. The introduced methods ensure detailed and scalable sentiment evaluations with applications in content personalization.

4. **Humanized AI Systems:** A programmable stance-directed AI architecture was proposed, enabling human-like, personalized social media content generation.

This framework aligns with user stances and preferences, fostering meaningful engagement and communication.

The findings of this thesis have significant implications for academia and industry:

- The integration of metadata-driven classification methods with sentiment and topic modeling paves the way for scalable AI solutions in digital marketing and social media analytics.
- Human-centered AI systems developed in this research offer practical tools for businesses seeking to enhance audience engagement and improve customer experiences through personalization.
- By addressing biases in sentiment analysis, this study contributes to creating fair and inclusive AI systems that uphold ethical principles in content generation.

While this thesis has made substantial contributions, several avenues for future work remain:

- Extending the proposed methods to other social media platforms to validate their generalizability and effectiveness in diverse contexts.
- Exploring the application of multimodal data, such as images and videos, alongside textual data to enhance user classification and content generation.
- Investigating the impact of real-time user feedback in refining personalized content and improving marketing strategies.
- Expanding the framework to address additional challenges in AI ethics, such as privacy concerns and transparency in automated decision-making processes.

This thesis demonstrates AI's transformative potential in bridging the gap between social media insights and actionable marketing strategies. This research contributes to developing human-centered AI systems that empower businesses and enhance user

experiences by prioritizing personalization, coherence, and fairness. The methodologies and findings presented here provide a foundation for continued innovation at the intersection of AI, social media, and targeted marketing.

REFERENCES

[1] L. De Silva and E. Riloff, "User type classification of tweets with implications for event recognition," in *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pp. 98–108, 2014.

[2] J. McCorriston, D. Jurgens, and D. Ruths, "Organizations are users too: Characterizing and detecting the presence of organizations on twitter," in *Ninth International AAAI Conference on Web and Social Media*, 2015.

[3] Z. Wood-Doughty, P. Mahajan, and M. Dredze, "Johns hopkins or johnny-hopkins: Classifying individuals versus organizations on twitter," in *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pp. 56–61, 2018.

[4] I. Samborskii, A. Filchenkov, G. Korneev, and A. Farseev, "Person, organization, or personage: Towards user account type prediction in microblogs," in *International Conference on Electronic Governance and Open Society: Challenges in Eurasia*, pp. 111–122, Springer, 2018.

[5] K. E. Daouadi, R. Z. Rebaï, and I. Amous, "Organization vs. individual: Twitter user classification," in *Proceedings of the second Conference on Language Processing and Knowledge Management*, vol. 2279, 2018.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[7] G. Tavares and A. Faisal, "Scaling-laws of human broadcast communication enable distinction between human, corporate and robot twitter users," *PloS one*, vol. 8, no. 7, 2013.

[8] G. Tavares, S. Mastelini, *et al.*, "User classification on online social networks

by post frequency," in *Anais do XIII Simpósio Brasileiro de Sistemas de Informação*, pp. 464–471, SBC, 2017.

[9] S. M. Kim, C. Paris, R. Power, and S. Wan, "Distinguishing individuals from organisations on twitter," in *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 805–806, 2017.

[10] S. Alzahrani, C. Gore, A. Salehi, and H. Davulcu, "Finding organizational accounts based on structural and behavioral factors on twitter," in *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pp. 164–175, Springer, 2018.

[11] Z. Wang, S. A. Hale, D. I. Adelani, P. A. Grabowicz, T. Hartmann, F. Flöck, and D. Jurgens, "Demographic inference and representative population estimates from multilingual social media data," in *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019* (L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, and L. Zia, eds.), pp. 2056–2067, ACM, 2019.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2261–2269, IEEE Computer Society, 2017.

[14] M. R and B. A, "Stance detection and mobile app recommendation discourse on tweets," *Comput. Intell.*, vol. 35, no. 4, pp. 1042–1059, 2019.

[15] D. Karatay and P. Karagoz, "User interest modeling in twitter with named entity recognition," in *Proceedings of the the 5th Workshop on Making Sense of Microposts co-located with the 24th International World Wide Web Conference (WWW 2015), Florence, Italy, May 18th, 2015*, vol. 1395 of *CEUR Workshop Proceedings*, pp. 17–20, 2015.

[16] A. Stavrianou and C. Brun, "Expert recommendations based on opinion mining

of user-generated product reviews," *Comput. Intell.*, vol. 31, no. 1, pp. 165–183, 2015.

[17] E. Tutaysalgir, P. Karagoz, and I. H. Toroslu, "Clustering based personality prediction on turkish tweets," in *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 825–828, IEEE, 2019.

[18] O. Bayat, O. N. Ucan, *et al.*, "Estimation of twitter user's nationality based on friends and followers information," *Computers & Electrical Engineering*, vol. 66, pp. 517–530, 2018.

[19] S. R. Sahoo and B. B. Gupta, "Hybrid approach for detection of malicious profiles in twitter," *Computers & Electrical Engineering*, vol. 76, pp. 65–81, 2019.

[20] E. Colleoni, A. Rozza, and A. Arvidsson, "Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data," *Journal of communication*, vol. 64, no. 2, pp. 317–332, 2014.

[21] Y. M. Çetinkaya, İ. H. Toroslu, and H. Davulcu, "Developing a twitter bot that can join a discussion using state-of-the-art architectures," *Social Network Analysis and Mining*, vol. 10, no. 1, pp. 1–21, 2020.

[22] R. Napoli, A. M. Ertugrul, A. Bozzon, and M. Brambilla, "A user modeling pipeline for studying polarized political events in social media," in *International Conference on Web Engineering*, pp. 101–114, Springer, 2018.

[23] J. S. Alowibdi, U. A. Buy, and P. Yu, "Language independent gender classification on twitter," in *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*, pp. 739–743, 2013.

[24] M. Vicente, J. P. Carvalho, and F. Batista, "Using unstructured profile information for gender classification of portuguese and english twitter users," in *International Symposium on Languages, Applications and Technologies*, pp. 57–64, Springer, 2015.

[25] I. Ha, K.-J. Oh, and G.-S. Jo, "Personalized advertisement system using social relationship based user modeling," *Multimedia Tools and Applications*, vol. 74, no. 20, pp. 8801–8819, 2015.

[26] T. P. Tran, "Personalized ads on facebook: An effective marketing tool for online marketers," *Journal of Retailing and Consumer Services*, vol. 39, pp. 230–242, 2017.

[27] F. De Keyzer, N. Dens, and P. De Pelsmacker, "How and when personalized advertising leads to brand attitude, click, and wom intention," *Journal of Advertising*, pp. 1–18, 2021.

[28] A. Simsek and P. Karagoz, "Wikipedia enriched advertisement recommendation for microblogs by using sentiment enhanced user profiles," *Journal of Intelligent Information Systems*, vol. 54, no. 2, pp. 245–269, 2020.

[29] I. Kangas, M. Schwoerer, and L. J. Bernardi, "Recommender systems for personalized user experience: lessons learned at booking.com," in *Proceedings of the 15th ACM Conference on Recommender Systems*, pp. 583–586, 2021.

[30] P. Mitsoulis-Ntompos, D. Varelas, T. Brady, J. E. Landry, R. F. Dickerson, T. Renner, C. Harris, S. Ye, A. Amirabadi, L. Jones, *et al.*, "Landing page personalization at expedia group," in *Proceedings of the 1st International Workshop on Industrial Recommendation Systems at KDD*, vol. 20, 2020.

[31] J. Wang, Z. Hao, M. Zhou, J. Chen, H. Zhou, Z. Song, J. Wang, J. Yang, and S. Ni, "Generating personalized titles incorporating advertisement profile," in *Database Systems for Advanced Applications: 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11–14, 2021, Proceedings, Part III 26*, pp. 529–540, Springer, 2021.

[32] T. Semerádová and P. Weinlich, *Website quality and shopping behavior: Quantitative and qualitative evidence*. Springer Nature, 2020.

[33] T. J. Cahill, B. Wertz, Q. Zhong, A. Parlato, J. Donegan, R. Forman, S. Manot, T. Wu, Y. Xu, J. J. Cummings, *et al.*, "The search for consumers of web-based raw dna interpretation services: using social media to target hard-to-reach pop-

ulations," *Journal of Medical Internet Research*, vol. 21, no. 7, p. e12980, 2019.

[34] C. Toraman, F. Şahinuç, E. H. Yilmaz, and I. B. Akkaya, "Understanding social engagements: A comparative analysis of user and text features in twitter," *Social network analysis and mining*, vol. 12, no. 1, p. 47, 2022.

[35] R. Barzilay and N. Elhadad, "Inferring strategies for sentence ordering in multidocument news summarization," *Journal of Artificial Intelligence Research*, vol. 17, pp. 35–55, 2002.

[36] D. Bollegala, N. Okazaki, and M. Ishizuka, "A bottom-up approach to sentence ordering for multi-document summarization," *Information processing & management*, vol. 46, no. 1, pp. 89–109, 2010.

[37] Y. Yin, L. Song, J. Su, J. Zeng, C. Zhou, and J. Luo, "Graph-based neural sentence ordering," *arXiv preprint arXiv:1912.07225*, 2019.

[38] B. Cui, Y. Li, and Z. Zhang, "Bert-enhanced relational sentence ordering network," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6310–6320, 2020.

[39] Z. Shao, M. Huang, J. Wen, W. Xu, and X. Zhu, "Long and diverse text generation with planning-based hierarchical variational model," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3257–3268, 2019.

[40] W. Yu, C. Zhu, T. Zhao, Z. Guo, and M. Jiang, "Sentence-permuted paragraph generation," *arXiv preprint arXiv:2104.07228*, 2021.

[41] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multim. Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, 2019.

[42] S. Sia, A. Dalmia, and S. J. Mielke, "Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too!," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1728–1736, 2020.

[43] 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.

[44] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[45] J. Jagarlamudi, H. D. III, and R. Udupa, "Incorporating lexical priors into topic models," in *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012* (W. Daelemans, M. Lapata, and L. Màrquez, eds.), pp. 204–213, The Association for Computer Linguistics, 2012.

[46] R. Egger and J. Yu, "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts," *Frontiers in sociology*, vol. 7, 2022.

[47] X. Li, L. Bing, W. Zhang, and W. Lam, "Exploiting BERT for end-to-end aspect-based sentiment analysis," in *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*.

[48] Y. Mao, Y. Shen, C. Yu, and L. Cai, "A joint training dual-mrc framework for aspect based sentiment analysis," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, pp. 13543–13551, 2021.

[49] L. Xu, Y. K. Chia, and L. Bing, "Learning span-level interactions for aspect sentiment triplet extraction," *arXiv preprint arXiv:2107.12214*, 2021.

[50] H. Yan, J. Dai, X. Qiu, Z. Zhang, *et al.*, "A unified generative framework for aspect-based sentiment analysis," *arXiv preprint arXiv:2106.04300*, 2021.

[51] C. D. Hoang, Q. V. Dinh, and N. H. Tran, "Aspect-category-opinion-sentiment extraction using generative transformer model," in *2022 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pp. 1–6, 2022.

[52] K. Scaria, H. Gupta, S. Goyal, S. A. Sawant, S. Mishra, and C. Baral, "Instructabsa: Instruction learning for aspect based sentiment analysis," *arXiv preprint arXiv:2302.08624*, 2023.

[53] X. Mei, Y. Zhou, C. Zhu, M. Wu, M. Li, and S. Pan, "A disentangled linguistic graph model for explainable aspect-based sentiment analysis," *Knowledge-Based Systems*, vol. 260, p. 110150, 2023.

[54] R. M. Marcacini and E. Silva, "Aspect-based sentiment analysis using bert with disentangled attention," *LatinX in AI at International Conference on Machine Learning 2021*, 2021.

[55] E. Park, R. K. Wong, and J. Kwon, "Can ai outperform human experts in creating social media creatives?," *arXiv preprint arXiv:2404.00018*, 2024.

[56] A. Mukherjee, "Safeguarding marketing research: The generation, identification, and mitigation of ai-fabricated disinformation," *arXiv preprint arXiv:2403.14706*, 2024.

[57] J. Ricker, D. Assenmacher, T. Holz, A. Fischer, and E. Quiring, "Ai-generated faces in the real world: a large-scale case study of twitter profile images," in *Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses*, pp. 513–530, 2024.

[58] M. Ienca, "On artificial intelligence and manipulation," *Topoi*, vol. 42, no. 3, pp. 833–842, 2023.

[59] R. Vinay, G. Spitale, N. Biller-Andorno, and F. Germani, "Emotional manipulation through prompt engineering amplifies disinformation generation in ai large language models," *arXiv preprint arXiv:2403.03550*, 2024.

[60] M. Carroll, A. Chan, H. Ashton, and D. Krueger, "Characterizing manipulation from ai systems," in *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–13, 2023.

[61] E. Häglund and J. Björklund, "Ai-driven contextual advertising: A technology report and implication analysis," *arXiv preprint arXiv:2205.00911*, 2022.

[62] Y. Wei and G. Tyson, "Understanding the impact of ai-generated content on social media: The pixiv case," in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6813–6822, 2024.

[63] C. Tarsney, "Deception and manipulation in generative ai," *arXiv preprint arXiv:2401.11335*, 2024.

[64] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[65] T. C. Helmus, "Artificial intelligence, deepfakes, and disinformation," *RAND Corporation*, pp. 1–24, 2022.

[66] C. Xu, Z. He, Z. He, and J. McAuley, "Leashing the inner demons: Self-detoxification for language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 11530–11537, 2022.

[67] R. Siegelmann, N. Mehrabi, P. Goyal, P. Goyal, L. Bauer, J. Dhamala, A. Galstyan, R. Gupta, and R. Ghanadan, "Mico: Preventative detoxification of large language models through inhibition control," in *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1696–1703, 2024.

[68] W. Li, W. Wei, K. Xu, W. Xie, D. Chen, and Y. Cheng, "Reinforcement learning with token-level feedback for controllable text generation," *arXiv preprint arXiv:2403.11558*, 2024.

[69] Y. M. Çetinkaya, M. Gürlek, I. H. Toroslu, and P. Karagöz, "Twitter account classification using account metadata: Organization vs. individual," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 30, no. 4, pp. 1404–1418, 2022.

[70] M. Mohd, R. Jan, and N. Hakak, "Enhanced bootstrapping algorithm for automatic annotation of tweets," *Int. J. Cogn. Informatics Nat. Intell.*, vol. 14, no. 2, pp. 35–60, 2020.

[71] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?," *IEEE Trans. Dependable Secur. Comput.*, vol. 9, no. 6, pp. 811–824, 2012.

[72] L. S. Martinez, M.-H. Tsou, and B. H. Spitzberg, "A case study in belief surveillance, sentiment analysis, and identification of informational targets for

e-cigarettes interventions," in *Proceedings of the 10th International Conference on Social Media and Society*, pp. 15–23, 2019.

[73] C. Lu, W. Lam, and Y. Zhang, "Twitter user modeling and tweets recommendation based on wikipedia concept graph," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[74] A. García-Silva, V. Rodríguez-Doncel, and Ó. Corcho, "Semantic characterization of tweets using topic models: A use case in the entertainment domain," *Int. J. Semantic Web Inf. Syst.*, vol. 9, no. 3, pp. 1–13, 2013.

[75] S. K. Bharti, R. Pradhan, K. S. Babu, and S. K. Jena, "Sarcastic sentiment detection based on types of sarcasm occurring in twitter data," *Int. J. Semantic Web Inf. Syst.*, vol. 13, no. 4, pp. 89–108, 2017.

[76] N. V. Patel and H. Chhinkaniwala, "Investigating machine learning techniques for user sentiment analysis," *Int. J. Decis. Support Syst. Technol.*, vol. 11, no. 3, pp. 1–12, 2019.

[77] J. R. Alharbi and W. S. Alhalabi, "Hybrid approach for sentiment analysis of twitter posts using a dictionary-based approach and fuzzy logic methods: Study case on cloud service providers," *Int. J. Semantic Web Inf. Syst.*, vol. 16, no. 1, pp. 116–145, 2020.

[78] F. Atefeh and W. Khreich, "A survey of techniques for event detection in twitter," *Computational Intelligence*, vol. 31, no. 1, pp. 132–164, 2015.

[79] D. Ramachandran and R. Parvathi, "Enhanced event detection in twitter through feature analysis," *International Journal of Information Technology and Web Engineering (IJITWE)*, vol. 14, no. 3, pp. 1–15, 2019.

[80] R. Srivastava and M. P. S. Bhatia, "Real-time unspecified major sub-events detection in the twitter data stream that cause the change in the sentiment score of the targeted event," *International Journal of Information Technology and Web Engineering (IJITWE)*, vol. 12, no. 4, pp. 1–21, 2017.

[81] C. Toraman, "Early prediction of public reactions to news events using microblogs," in *Seventh BCS-IRSG Symposium on Future Directions in Information Access 7*, pp. 1–4, 2017.

[82] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, 2016.

[83] F. Poldi, "Twint-twitter intelligence tool," *URL: https://github.com/twintproject/twint (visited on 01/27/2020)*, 2019.

[84] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[85] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Association for Computational Linguistics, 2018.

[86] Y. Shibata, T. Kida, S. Fukamachi, M. Takeda, A. Shinohara, T. Shinohara, and S. Arikawa, "Byte pair encoding: A text compression scheme that accelerates pattern matching," tech. rep., Technical Report DOI-TR-161, Department of Informatics, Kyushu University, 1999.

[87] B. Heinzerling and M. Strube, "BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), 2018.

[88] 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, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[89] R. Ferguson, "Word of mouth and viral marketing: taking the temperature of the hottest trends in marketing," *Journal of consumer marketing*, 2008.

[90] Y. M. Çetinkaya, E. Külah, İ. H. Toroslu, and H. Davulcu, "Targeted marketing on social media: utilizing text analysis to create personalized landing pages," *Social Network Analysis and Mining*, vol. 14, no. 1, p. 77, 2024.

[91] T. Ash, *Landing page optimization: The definitive guide to testing and tuning for conversions*. John Wiley & Sons, 2011.

[92] Y. M. Çetinkaya, I. H. Toroslu, and H. Davulcu, "Coherent personalized paragraph generation for a successful landing page," in *Proceedings of the Advances in Social Network Aanalysis and Mining (ASONAM'22)*, (Istanbul, Turkey), 2022.

[93] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "Yake! keyword extraction from single documents using multiple local features," *Information Sciences*, vol. 509, pp. 257–289, 2020.

[94] N. Firoozeh, A. Nazarenko, F. Alizon, and B. Daille, "Keyword extraction: Issues and methods," *Natural Language Engineering*, vol. 26, no. 3, pp. 259–291, 2020.

[95] J. Liu, J. Shang, and J. Han, "Phrase mining from massive text and its applications," *Synthesis Lectures on Data Mining and Knowledge Discovery*, vol. 9, no. 1, pp. 1–89, 2017.

[96] 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.

[97] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[98] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[99] H. Zhu and K. Liu, "Temporal, spatial, and socioeconomic dynamics in social

media thematic emphases during typhoon mangkhut," *Sustainability*, vol. 13, no. 13, p. 7435, 2021.

[100] J. L. Fleiss, "Measuring nominal scale agreement among many raters.," *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[101] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[102] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, "Is chatgpt a general-purpose natural language processing task solver?," 2023.

[103] OpenAI, "Gpt-4 technical report," 2023.

[104] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023.

[105] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[106] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[107] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824–24837, 2022.

[108] Y. Lee, Y. M. Çetinkaya, E. Külah, İ. H. Toroslu, and H. Davulcu, "Masking the bias: From echo chambers to large scale aspect-based sentiment analysis," in *2024 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Springer, 2024.

[109] S. U. S. Chebolu, F. Dernoncourt, N. Lipka, and T. Solorio, "A review of datasets for aspect-based sentiment analysis," in *International Joint Conference on Natural Language Processing*, 2022.

[110] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, "Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship," in *Proceedings of the 2018 world wide web conference*, pp. 913–922, 2018.

[111] C. Torres-Lugo, K.-C. Yang, and F. Menczer, "The manufacture of partisan echo chambers by follow train abuse on twitter," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, pp. 1017–1028, 2022.

[112] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A survey on aspect-based sentiment analysis: Tasks, methods, and challenges," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[113] D. Cherepnalkoski and I. Mozetič, "Retweet networks of the european parliament: Evaluation of the community structure," *Applied network science*, vol. 1, pp. 1–20, 2016.

[114] O. C. Seckin, A. Atalay, E. Otenen, U. Duygu, and O. Varol, "Mechanisms driving online vaccine debate during the covid-19 pandemic," *Social Media+ Society*, vol. 10, no. 1, p. 20563051241229657, 2024.

[115] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[116] T. Elmas, M. Randl, and Y. Attia, "# teamfollowback: Detection & analysis of follow back accounts on social media," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 18, pp. 381–393, 2024.

[117] Y. Zhang, Y. Wu, and Q. Yang, "Community discovery in twitter based on user interests," *Journal of Computational Information Systems*, vol. 8, no. 3, pp. 991–1000, 2012.

[118] Y. Lee, M. Ozer, S. R. Corman, and H. Davulcu, "Identifying behavioral factors leading to differential polarization effects of adversarial botnets," *ACM SIGAPP Applied Computing Review*, vol. 23, no. 2, pp. 44–56, 2023.

[119] A. Rao, F. Morstatter, and K. Lerman, "Retweets amplify the echo chamber effect," in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pp. 30–37, 2023.

[120] S. Tsugawa and H. Ohsaki, "Negative messages spread rapidly and widely on social media," in *Proceedings of the 2015 ACM on conference on online social networks*, pp. 151–160, 2015.

[121] M. Mousavi, H. Davulcu, M. Ahmadi, R. Axelrod, R. Davis, and S. Atran, "Effective messaging on social media: What makes online content go viral?," in *Proceedings of the ACM Web Conference 2022*, pp. 2957–2966, 2022.

[122] L. Moustakas, "Social cohesion: Definitions, causes and consequences," *Encyclopedia*, vol. 3, no. 3, pp. 1028–1037, 2023.

[123] Y. M. Çetinkaya, Y. Lee, E. Külah, İ. H. Toroslu, M. A. Cowan, and H. Davulcu, " Toward a Programmable Humanizing Artificial Intelligence Through Scalable Stance-Directed Architecture ," *IEEE Internet Computing*, vol. 28, pp. 20–27, Sept. 2024.

[124] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27730–27744, 2022.

[125] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in *2010 43rd Hawaii international conference on system sciences*, pp. 1–10, IEEE, 2010.

[126] M. D. Buhrmester, M. A. Cowan, and H. Whitehouse, "What motivates barrier-crossing leadership?," *New England Journal of Public Policy*, vol. 34, no. 2, p. 7, 2022.

[127] N. Dakiche, F. B.-S. Tayeb, Y. Slimani, and K. Benatchba, "Tracking community evolution in social networks: A survey," *Information Processing & Management*, vol. 56, no. 3, pp. 1084–1102, 2019.

[128] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[129] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, A. Ashok, A. S. Dhanasekaran, A. Naik, D. Stap, *et al.*, "Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks," *arXiv preprint arXiv:2204.07705*, 2022.

[130] C. Toraman, F. Şahinuç, and E. H. Yilmaz, "Large-scale hate speech detection with cross-domain transfer," in *Proc. Language Resources and Evaluation Conf.*, p. 2215–2225, 2022.

CURRICULUM VITAE

PERSONAL

NAME : Yusuf Mücahit Çetinkaya

PLACE : Ankara, Türkiye

WORK EXPERIENCE

2017-	**Research and Teaching Assistant** Department of Computer Enginering, METU, Ankara, Türkiye
2023-2024	**Research Scholar** School of Computing and AI, Arizona State University, AZ, USA
2017-2021	**Technical Consultant** HAVELSAN, Ankara, Türkiye
2016-2017	**Software Engineer** HAVELSAN, Ankara, Türkiye
2013-2016	**Student Teaching Assistant** Department of Computer Enginering, METU, Ankara, Türkiye
2015	**Undergraduate Researcher** UMONS, Mons, Belgium

EDUCATION

2024 | Ph.D. in COMPUTER ENGINEERING
Graduate School of Natural and Applied Sciences
Middle East Technical University (METU), Türkiye
Thesis: "Bridging AI and Personalization: From Social Media Insights to Targeted Marketing"
Advisor: İsmail Hakkı TOROSLU

2019 | MSc. Degree in COMPUTER ENGINEERING
Graduate School of Natural and Applied Sciences
Middle East Technical University (METU), Türkiye
Thesis: "Developing a Twitter Bot that can Join a Discussion Using State-of-the-art Architectures"
Advisor: İsmail Hakkı TOROSLU

2016 | Bs Degree in COMPUTER ENGINEERING
Faculty of Engineering
Middle East Technical University (METU), Türkiye

PUBLICATIONS

Journal and conference publications are annotated with the "**J**" and "**C**" prefixes.

J **Çetinkaya, Y. M.**, Lee, Y., Külah, E., Toroslu, İ. H., Cowan, M. A., & Davulcu, H. (2024). Towards a Programmable Humanizing AI through Scalable Stance-Directed Architecture. IEEE Internet Computing, vol. 28, no. 05, pp. 20-27, doi: 10.1109/MIC.2024.3450090.

C Lee, Y., **Çetinkaya, Y. M.**, Külah, E., Toroslu, İ. H., & Davulcu, H. (2024). Masking the Bias: From Echo Chambers to Large Scale Aspect-Based Sentiment Analysis. In 2024 International Conference on Advances in Social Networks Analysis and Mining. Springer.

J **Çetinkaya, Y. M.**, Külah, E., Toroslu, İ. H., & Davulcu, H. (2024). Targeted marketing on social media: utilizing text analysis to create personalized landing pages. Social Network Analysis and Mining, 14(1), 77.

J Külah, E., **Çetinkaya, Y. M.**, Özer, A. G., & Alemdar, H. (2023). COVID-19

forecasting using shifted Gaussian Mixture Model with similarity-based estimation. Expert Systems with Applications, 214, 119034.

C Çetinkaya, Y. M., Toroslu, İ. H., & Davulcu, H. (2022). Coherent personalized paragraph generation for a successful landing page. In 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 252-255). IEEE.

J Cetinkaya, Y. M., Gürlek, M., Toroslu, I. H., & Karagöz, P. (2022). Twitter account classification using account metadata: organization vs. individual. Turkish Journal of Electrical Engineering and Computer Sciences, 30(4), 1404-1418.

J Çetinkaya, Y. M., Toroslu, İ. H., & Davulcu, H. (2020). Developing a Twitter bot that can join a discussion using state-of-the-art architectures. Social network analysis and mining, 10, 1-21.

C Çetinkaya, Y. M., Doğu Z. S., & Dayıbaş O. (2017). On the Development of the Autonomous City. In 2018 International Symposium on Innovative Technologies in Engineering and Science (ISITES) (pp. 407-417).