Chapter 1

Introduction

1.1 What are Social Media?

Social media (SM) encompasses a diverse array of online platforms designed to facilitate connectivity, information dissemination, and interaction among users across the globe. It serves as a virtual meeting place where individuals converge to share thoughts, ideas, and multimedia content, akin to an expansive digital forum fostering an exchange of perspectives and experiences. At its core, social media is characterized by several salient features:

- **Sharing:** Users can disseminate information in various content formats, including text, images, videos, and hyperlinks. This functionality empowers individuals to express themselves creatively and broadcast their narratives to a larger audience than otherwise physically possible.
- **Interaction:** Social media platforms cultivate an interactive milieu wherein users can engage with each other's posts through likes, comments, and shares. These mechanisms facilitate dialogue, engender community cohesion, and fortify interpersonal relationships in the digital realm.
- **Communities:** A fundamental facet of social media is its capacity to forge connections among individuals with shared interests, affiliations, or demographics.

By fostering the formation of virtual communities, these platforms facilitate the cultivation of interpersonal bonds, enabling users to connect with friends, family, acquaintances, and like-minded individuals worldwide.

- **Content Creation:** Social media democratizes the process of content creation, empowering users of all backgrounds and skill levels to generate and publish their material. This democratization stitches together a rich tapestry of diverse content, from amateur endeavors to polished productions, enriching the collective digital landscape.

In addition to the above attributes, social media platforms have evolved to provide myriad functionalities and services, catering to an expansive array of user preferences and objectives. These platforms serve as conduits for news dissemination, educational resources, entertainment, e-commerce transactions, and civic engagement, among other pursuits. Prominent exemplars of social media include Facebook, Instagram, Twitter, YouTube, and TikTok, each boasting vast user bases and wielding significant cultural influence (see Figure 1.1). The percentage points show the market share of individual platforms in social media considering mobile-only devices and all devices respectively. The ubiquity and omnipresence of social media in contemporary society underscore its profound impact on various facets of human interaction, communication, and culture. As such, understanding the dynamics and implications of social media has emerged as a paramount area of scholarly inquiry, encompassing disciplines ranging from sociology and psychology to communication studies and digital humanities. Social media epitomizes a transformative force in the digital age, redefining the dynamics of human connectivity, expression, and community engagement. Its multifaceted nature and far-reaching implications underscore its significance as both a technological artifact and a socio-cultural phenomenon warranting scholarly scrutiny and discourse.

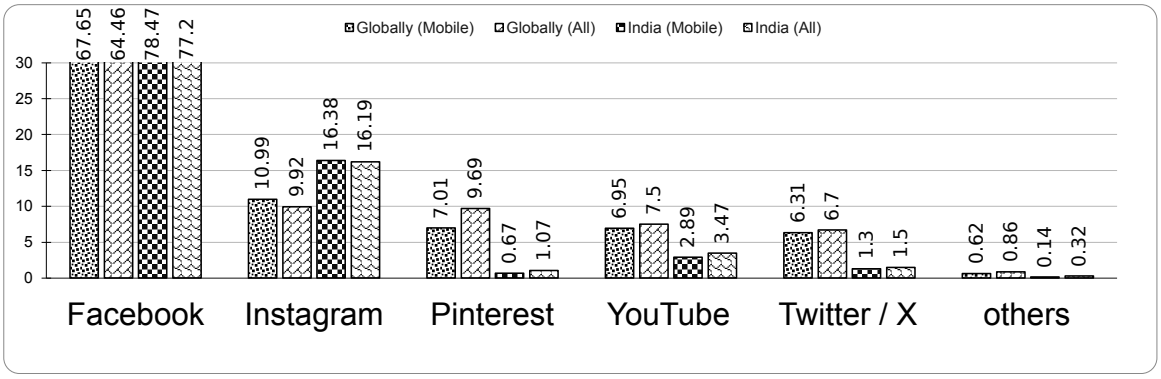


Figure 1.1: Market shares of leading SM platforms (in percentages)

1.2 Rise of Social Media

The ascension of social media unfolds as a narrative encapsulating two pivotal revolutions: the proliferation of the internet and the advent of mobile technology. The nascent stages of this phenomenon can be traced back to the 1980s and 90s, characterized by the emergence of rudimentary online communities revolving around email, bulletin boards, and chatrooms. However, these early endeavors were marked by their cumbersome nature and constrained functionalities in comparison to the subsequent evolution. The turn of the millennium witnessed a paradigm shift with the emergence of social media platforms such as Six Degrees (1997) and Friendster (2002), which swiftly gained prominence [2]. Notably, MySpace attained a milestone in 2003 by surpassing one million monthly active users, facilitating the creation of user profiles and fostering online connections among individuals. The first smartphone to enter the market was IBM's Simon Personal Communicator (SPC), which made its debut in 1994 [1]. Subsequently, the first iPhone was introduced in 2007, while Android unveiled the HTC Dream phone in 2008. This marked the onset of a transformative era, signifying the migration of social media usage from desktop platforms to the pervasive domain of handheld devices [2]. Capitalizing on this paradigm shift, Facebook emerged as a frontrunner in the

[1] `https://www.bbc.com/news/technology-28802053`, retrieved 5 Apr 2024

[2] `https://gs.statcounter.com/social-media-stats/mobile/worldwide/2009`, retrieved 5 Apr 2024

late 2000s, bolstered by its mobile-friendly interface. Subsequently, an array of social media behemoths including Twitter, Instagram, and YouTube emerged, each catering to distinct modes of content dissemination. The meteoric rise of social media is underscored by its exponential growth trajectory. As of January 2024, global internet users number 5.35 billion, constituting 66.2 percent of the world's populace, with 5.04 billion individuals, or 62.3 percent of the global population, engaging with social media platforms. In the Indian context, the prevalence of active social media users amount to approximately 470.1 million as of 2022 (see Figure 1.2 [3]), indicating its pervasive integration into societal fabric.

Social media platforms have emerged as rapid sources of information dissemination and consumption pertaining to both local events and global news. These platforms facilitate instantaneous sharing of information and communication among individuals regardless of geographical location. A study conducted by Oxford University reveals that 54% of Indians rely on social media platforms for accessing "truthful" information, surpassing the global average of 37% and the corresponding figure in the United States, which stands at 29%. However, SM contents are mostly generated by the users, unlike the traditional information sharing platforms like websites of trusted sources where professionals are employed. These contents generated by users, or user-generated content (UGC) predominantly exhibit informal characteristics, often deviating from the grammatical rules and standard formats typical of formal communication. Consequently, social media text is marked by non-standard spellings of words, self-generated abbreviations, and non-standard grammatical structures. Additionally, for many languages, native keyboards are not readily available on computing devices. Even when available, a significant portion of users may not be proficient in using their native language keyboards or may opt for phonetic typing (transliterated or Romanized text) and code-mixing (CM) for convenience. The SM text is thus often multi-lingual and multi-script.

[3]Pic Courtesy: `https://localiq.com/blog/what-happens-in-an-internet-minute/`, retrieved 23 Mar 2024

Such linguistic diversity throws a new set of challenges that are specific to the SM text analysis.

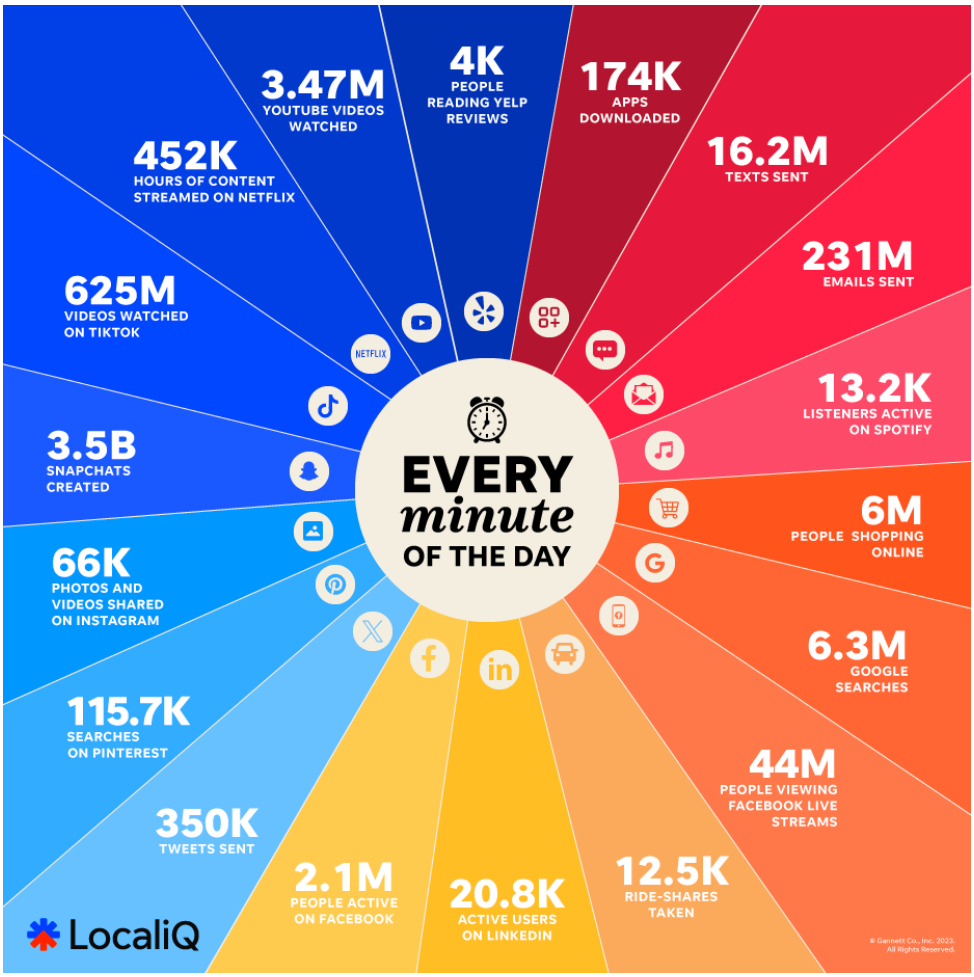


Figure 1.2: Social media infographics

1.3 Code-Mixing

Code-mixing (CM) is a phenomenon typical to informal communication and widely prevalent in social media text, in particular. It refers to occurrence of lexical items and grammatical features from two or more languages in a single sentence. Myers [3] defines code-mixing as an "embedding of linguistic units such as phrases, words, and morphemes of one language into an utterance of another language". Code-mixing hap-

pens all over the world with numerous languages. A few of the code-mixed languages are termed Spanglish (mixing of Spanish - English), Hinglish (mixing of Hindi-English), Tenglish (mixing of Telugu - English), Portunol (mixing of Portugese - English) and Franglais (mixing of French - English) [4]. Globally, approxiately 3.5% of all tweets are CM [5] and in India, 17.2% of all posts and comments are CM [6]. Often code-mixing is coupled with another phenomenon called transliteration. Transliteration is the process of phonetically converting words of a language into a foreign or non-native script. Due to a lack of technological support, people often write in a language with a non-native script. E.g., a lot of Indian writers express themselves in their native languages using the Roman script on electronic platforms. This phenomenon is called transliteration, and the text is transliterated text. Figure 1.3 displays a CM conversation taken from a Bollywood movie 'Pink'.

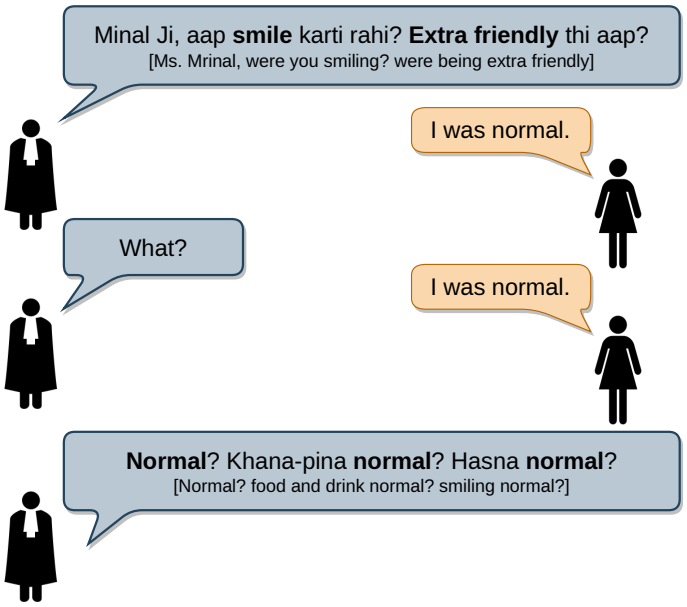
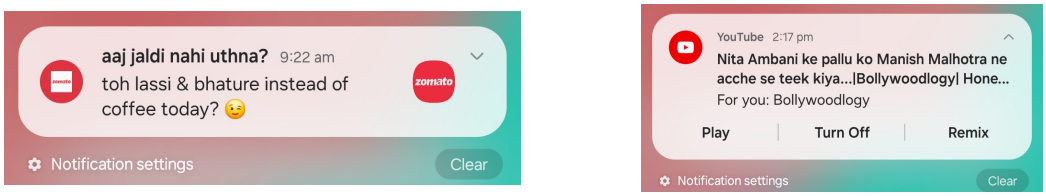


Figure 1.3: Example of a CM conversation from movie Pink

1.4 Why CM text processing is important?

It is imperative to recognize the inherent proficiency of humans in language construction and the acquisition of novel communication methodologies. Our aptitude extends not only to human-to-human communication but also to the efficient development and utilization of language frameworks for human-machine interaction. For instance, while initial interactions with machines involved the issuance of command-line expressions to execute specific tasks, now interactions with platforms such as Alexa emulate natural human-to-human conversations. Should we aspire for machines to engage in meaningful human-like dialogues, it is necessary that they possess the capability to comprehend and respond appropriately to the linguistic nuances embedded within such exchanges.

For corporate entities, proficiency in understanding code-switched communication holds considerable implications for enhanced advertisement targeting. Figures 1.4 exemplify instances where code-mixed language is employed to engage clients effectively. Discerning genuine user sentiment regarding product attributes facilitates the refinement of subsequent iterations. Agarwal et. al [7] identified a correlation between language and sentiment within a sentence, indicating that neglecting one language in favor of another influences the interpretation. Furthermore, disregarding code-switched languages entirely could potentially result in erroneous conclusions regarding user sentiment.



(a) Notification from Zomato

(b) Notification from YouTube

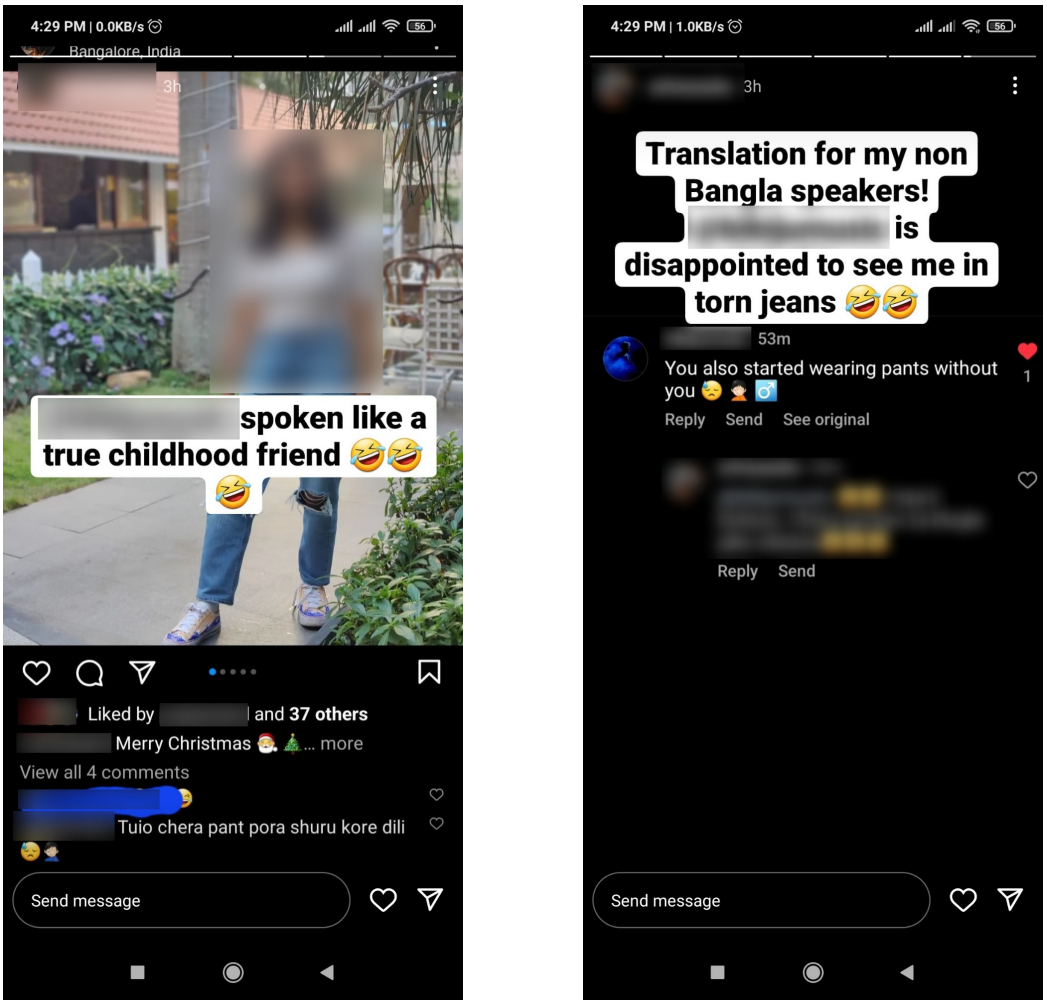
Figure 1.4: Example of CM notification push by companies

In healthcare, it is found that attaining insights into individuals' emotional states and receptiveness facilitates improve care delivery, enhanced patient communication,

and heightened uptake of preventive measures [8].

Statistics from the Journal of Neurolinguistics [9] indicate that 43% of the global population is bilingual, with 17% being multilingual or fluent in more than two languages. These people use multiple languages in informal communication, oral or written. Given this prevalence of multilingualism, a lot SM data is code-mixed. It is thus imperative to develop Natural Language Processing (NLP) technologies capable of effectively processing code-mixed (CM) data. The proliferation of social media platforms has catalysed the increase in code-mixed text. Even though recent years have witnessed a surge in CM-related research, it is no-where near desired level. For example, Figure 1.5 illustrates a scenario where machine translation systems falter in comprehending CM language, underscoring the necessity of further advancements in text processing technologies.

The rapid expansion of voice-operated devices has democratized access to smart assistants, resulting in interactions with NLP technologies by both monolingual and multilingual users. Winata et. al [1] conducted the first large-scale comprehensive survey on Code-Switching (CS) Natural Language Processing (NLP) research in a structured manner, amassing over 400 papers from open repositories including the ACL Anthology and ISCA proceedings. The entirety of the ACL Anthology repository up to October 2022 was systematically crawled. Subsequently, papers were filtered based on keywords pertinent to Code-Switching, namely “codeswitch”, “code switch”, etc. The top figure in Figure 1.6 represents the relative distribution compared to all *CL and ISCA papers, while the bottom one presents the absolute number of publications categorized by conferences versus workshops. It is important to note that these figures do not consider papers published after the specified time frame. Additionally, it is pertinent to mention that the graphs exclude the number of publications released in journals and symposiums. Both the figures showcase rapid surge in the research involving code-mixed text processing.



(a) Original comment on a post

(b) Incorrect translation of the comment

Figure 1.5: Example of an original comment and its machine translation version

1.5 Tasks on Code-Mixed data

This section discusses the diverse array of Natural Language Processing (NLP) tasks pertinent to code-mixed text. These tasks can be categorically performed at different levels like at word level, sentence level, and across the sentences (as depicted in Figure 1.7. Each category is expounded upon individually within the context of its respective task. Following are some of the tasks carried out on CM data.

1. Language identification (LID)
2. Named entity recognition(NER)

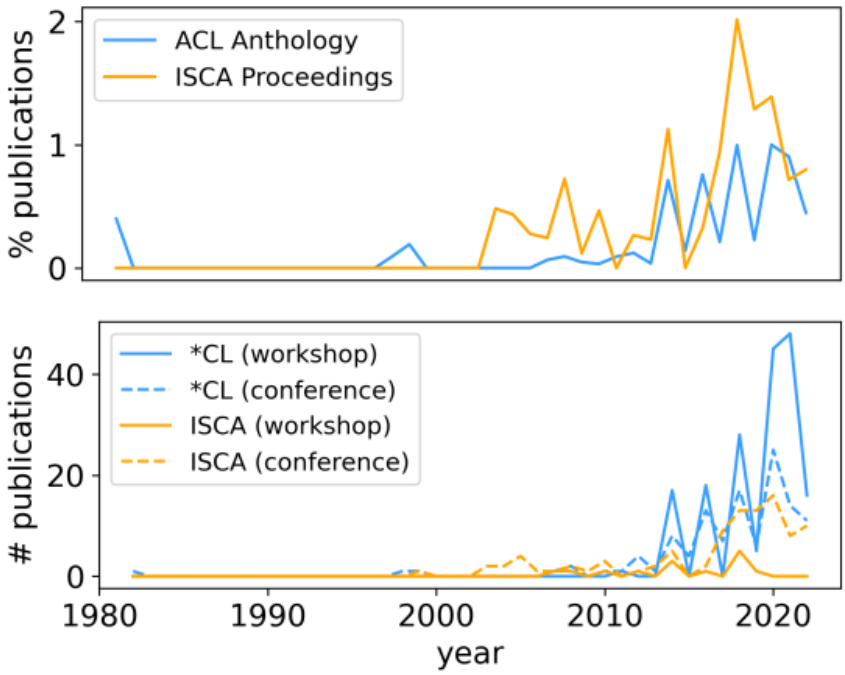


Figure 1.6: Number of publications over time in *CL and ISCA venues. [1]

3. Parts of speech (PoS)
4. Sentiment analysis (SA)
5. Hate speech and offensive content identification
6. Sarcasm detection
7. Question answering
8. Natural language inference
9. Information retrieval (IR)

Among the above, we focus on the following tasks in this thesis.

1.5.1 Language Identification (LID)

Language identification aims to identify the language affiliation of individual words within a text characterized by seamless integration of two or more languages. Specifically, our focus lies in devising a model capable of accurately assigning language tags to each word in a code-mixed sentence, as exemplified in Figure 1.8. In this example, *EN*

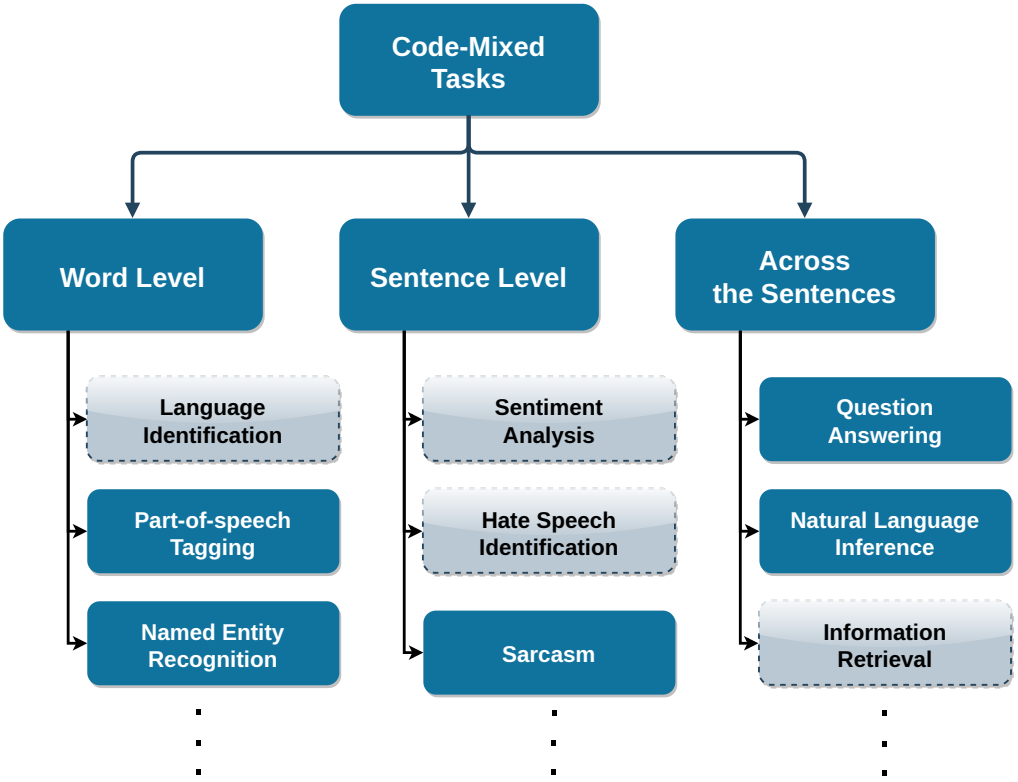


Figure 1.7: Type of NLP task on code-mixed data

stands for English and *HI* stands for Hindi. The meaning of the sentence is *Microsoft organised a worldwide Hackathon.*

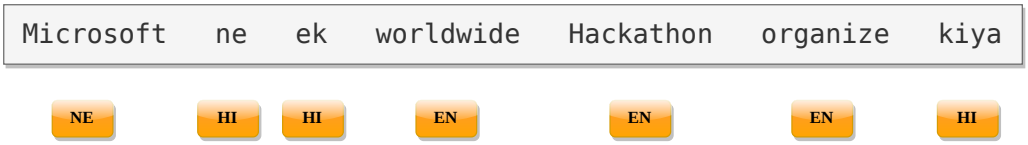


Figure 1.8: Example of language identification task

1.5.2 Sentiment Analysis

Sentiment analysis or opinion mining is a field in natural language processing which involves identifying and extracting subjective information from a text snippet. This is useful in various applications. For example, from customer feedback data, companies can use sentiment analysis to understand how customers feel about their products or

services and identify areas for improvement. In market research, sentiment analysis can help businesses track the opinions of customers or competitors in real-time and make informed decisions on marketing and product development. With social media monitoring, organizations can monitor the sentiment of their brand on social media platforms and fine tune their responses to negative or positive comments/feedback and calibrate their positions in the market and society. Governments and media organizations also can use sentiment analysis to track public mood on a specific issue or person. Sentiment analysis is thus important because it helps organizations understand how people feel about specific topics or products, which enables them to make informed business decisions and improve customer satisfaction. We focus here sentiment analysis on code-mixed data as social media data largely code-mixed.

1.5.3 Hate Speech and Offensive Content Identification

The domain of hate speech identification has undergone extensive scrutiny and investigation, with both academia and researchers dedicating substantial efforts on it in the recent time. The majority of this research has been conducted within a monolingual framework. Processing and analyzing code-mixed data is more challenging than its monolingual counterpart due to the presence of multiple languages and their different linguistic nuances. It is thus imperative for academicians and practitioners to come up with necessary tools and methodologies to effectively analyze and manage such data, particularly in the realm of hate speech identification, given the prevalence of code-mixed content on social media platforms. The objective here is to classify social media content into two distinct categories: hate and offensive (HOF) and non-offensive (NOT). Considering the structural aspect of hate speech or offensive content, they can be further refined into standalone hate (SHOF) and contextual hate (CHOF).

1.5.4 Information Retrieval

Information Retrieval is the process of searching relevant documents based on a user query that fulfills the user's information need. Understanding user need in itself is non-trivial for a machine, but it gets even more difficult when queries are code-mixed. The queries can be written in either a native script or Roman script which needs to be matched to the documents that can also be in either native script or Roman script or code-mixed. As described by Chakma and Das [10], when query terms and documents belong to different languages which may be using their native scripts or non-native ones, the retrieval set up is called code-mixed Information Retrieval (CMIR). Here, both query and documents can contain multiple languages and scripts. If $q \in \langle L^{(i)}, S^{(j)} \rangle$ where $i \geq 2$ and $j \geq 1$. where $L^(i)$ = union of i many languages and $S^(j)$ = union of j many scripts. Similarly, the document pool thus becomes

$$\mathcal{D} = \bigcup D_{L^{(i)},S^{(j)}}$$

where $L^{(i)} = \{l_1, l_2, \ldots, l_i\}$, $S^{(j)} = \{s_1, s_2, \ldots, s_j\}$ and

$D_{L^{(i)},S^{(j)}}$ = set of documents in language from $L^{(i)}$ written in script from $S^{(j)}$.

1.6 Motivation and Challenges

The exponential expansion of social media platforms and their intuitive interfaces have sparked our interest in delving into this realm. Our journey commenced with an exploration of social media data dynamics, leading us to the interesting domain of transliterated multilingual data. As we immersed ourselves in the analysis of data sourced from platforms like Facebook, WhatsApp, and others, the prevalence of code-mixed content became apparent. This mixing of languages permeates extensively across social media platforms, serving as a primary reason for our focus on code-mixed data.

One prominent observation from our exploration is that a large number of com-

munity groups are active on social media platforms such as Facebook and Whatsapp. These groups frequently engage in discussions, sharing information, and addressing problems in an informal conversational style, often employing a code-mixed language. This observation sparked our initial inquiry: how can we effectively retrieve the most relevant information amidst this linguistic amalgamation? This question inherently presents itself as an information retrieval challenge, wherein user queries serve as the query, and subsequent user replies function as the documents. However, the absence of suitable datasets prompted us to undertake the creation of a comprehensive dataset comprising a minimum of 50 queries and at least 100,000 documents.

Subsequently, we identified the task of language identification within code-mixed sentences or corpora as a fundamental one. While several models are available for the task, we find the efficacy of leveraging contextual embedding and transformer-based models to enhance performance in this domain. This observation also motivated us to delve into word-level language identification within code-mixed data.

We initially focused on collecting data of Bengali-English language pair for the information retrieval task. We then also considered Hindi-English language pairs. Given India's linguistic diversity, characterized by 22 official languages and numerous dialects, we further expanded to include three Dravidian language pairs: Tamil-English, Malayalam-English, and Kannada-English, for sentiment analysis task of YouTube comments.

Sentiment analysis is one of the principal applications within the natural language processing domain, motivated by the nuanced impact of code-mixing on sentiment expression. Conceptualized as a text classification task, sentiment analysis intrigued by code-mixed content posed unique challenges that we find as an interesting research area.

Similarly, the identification of hate speech represented another crucial text classification task within our purview. However, traditional classification models proved insufficient in effectively identifying conversational hate speech. Acquiring suitable

training data for model development was a major challenge. However, it was partially alleviated by the availability of datasets such as those created by the HASOC [4] team, particularly for exploration within the Hindi-English language pair.

In summary, text processing on code-mixed social media data is a relatively new area that demands research and innovation in developing newer tools and techniques because of its fast proliferation among the masses. While the scope is huge, it is handicapped by inadequate ready datasets, tools and techniques. Thus the domain has its innate motivation and research challenges that has driven us to the exciting field in this dissertation work.

1.7 Dissertation Overview

The primary objective of this dissertation is to investigate text processing techniques applied to code-mixed social media data. Our exploration encompasses various tasks crucial in this domain, including word-level language identification, sentiment analysis, hate speech detection, and code-mixed information retrieval. Furthermore, we employ a range of models and algorithms aimed at enhancing the effectiveness of social media applications across these specific areas.

Initially, we develop a deep learning-based word-level language identification system tailored for code-mixed data, addressing three language pairs: Bengali-English, Hindi-English, and Spanish-English.

Subsequently, acknowledging the linguistic diversity prevalent in India, we delve into sentiment analysis using three Dravidian code-mixed language datasets: Tamil-English, Kannada-English, and Malayalam-English, using the FIRE 2021 dataset. Our approach involves proposing a model that integrates a pre-trained model with word-level language tagging techniques.

We then work on the detection of hate speech and offensive content within code-

[4] https://hasocfire.github.io/hasoc/2021/index.html

mixed conversations on social media platforms. We utilize the HASOC 2021 dataset, comprising conversational tweets, comments, and replies structured hierarchically. Our proposed methodology involves fine-tuning the mBERT model with contrastive loss and subsequently constructing an ensemble model incorporating a sentence transformer.

Finally, we explore information retrieval techniques tailored for code-mixed data. This entails detailing the corpus collection, creation, and annotation processes. Additionally, we employ phonetic encoding for query expansion, diverse stop words removal techniques across languages to enhance information retrieval performance.

1.8 Research Goals

The primary aim of this research is to investigate text processing methodologies applied to code-mixed social media data, with a particular emphasis on the development of an enhanced system tailored for various downstream tasks. Specifically, our focus encompasses four pivotal downstream tasks: word-level language identification, sentiment analysis, identification of conversational hate speech, and information retrieval from code-mixed datasets. While each of these tasks is scrutinized with specific research inquiries described in separate chapters, the overarching endeavor is directed towards addressing the following research objectives (ROs) within our study.

- **RO1:** What pivotal roles does language identification serve in the context of sentiment analysis?
- **RO2:** What methodologies can be employed to discern hate speech and offensive content within conversational threads?
- **RO3:** In what capacity does language identification contribute to code-mixed information retrieval?
- **RO4:** How does language identification facilitate the removal of stopwords and improve code-mixed IR?

It is noteworthy that RO1, part of the Dravidian CodeMix shared task, was based

on data collected by the task organizers, while RO2 drew upon alternate code-mixed data obtained from the HASOC shared task organizers. Furthermore, RO3 and RO4 were explored through the analysis of social media posts sourced from the platform Facebook. Beyond the elucidation of these research questions, an additional imperative of our investigation entails the establishment of a standardized benchmark datasets for code-mixed information retrieval in Indian languages, thus underscoring the importance for ongoing scholarly inquiry in this domain.

1.9 Contribution

This dissertation presents several foundational contributions aimed at enhancing text processing in code-mixed social media data. Notably, it introduces a novel dataset designed for code-mixed information retrieval, while also employing a range of deep learning techniques on existing datasets to assess the efficacy of current systems across various tasks. We highlight below the main contributions of this dissertation.

- Primarily, this research addresses a fundamental task crucial for code-mixed datasets, namely language identification at the word level. A significant contribution lies in the comparison between non-contextual input representations (Word2Vec, GloVe, FastText) and contextual input representation (BERT) in automatically identifying languages within code-mixed data. The proposed approach harnesses a deep learning framework that capitalizes on pre-trained models for embedding, offering a pioneering solution to this challenge. Evaluation of the proposed model spans six distinct datasets, covering three language pairs: Bengali-English, Hindi-English, and Spanish-English. Additionally, the proportion of code-mixed data within each dataset is quantified, and the model's performance is evaluated under varying conditions, including scenarios featuring solely code-mixed data and combinations of code-mixed and monolingual data.
- Secondly, this study endeavors to ascertain the sentiment polarity of code-mixed

data extracted from Dravidian language pairs (Malayalam-English, Tamil-English, and Kannada-English), sourced from comments on YouTube videos obtained through social media channels. The text is categorized into five classes: Positive, Negative, mixed feelings, unknown state, and not in language. Pertinent inquiries include delineating criteria for labeling a dataset as code-mixed and identifying optimal strategies for sentiment analysis across datasets featuring diverse text compositions. The importance of language identification (LID) in code-mixed data processing is scrutinized, along with quantifying potential efficacy of employing a separate language identification model featuring pre-trained models to enhance overall system performance, particularly in the context of sentiment analysis. Furthermore, the feasibility of viewing sentiment analysis as a multi-level problem, as opposed to a multi-class problem, is explored, along with the potential benefits of employing a multi-level hierarchical model to augment performance.

- In the third contribution, this study confronts the challenge of identifying hate speech and offensive content within code-mixed conversations on social media platforms. Given the presence of multiple languages within a single conversation, this task is subdivided into two components: binary classification and multi-class classification. The former entails determining whether a tweet, comment, or reply contains hate speech, offensive language, or profanity (HOF), or is devoid of such content and categorized as non-hate and non-offensive (NOT). The latter involves identifying hate speech within code-mixed conversations by discerning specific forms of hate expression, including standalone hate (SHOF) and contextual hate (CHOF), which pertains to content supporting hate expressed in the parent conversation. Various deep learning models, including Bidirectional Encoder Representations from Transformers (BERT), are employed for classification purposes. Additionally, an ensemble model comprising a fine-tuned mBERT model and a sentence transformer is proposed to enhance classification accuracy.

- Finally, the fourth contribution delves into the challenges inherent in information retrieval from code-mixed data. Collecting raw data from social media platforms, followed by cleaning and pre-processing to make it ready and suitable for information retrieval experiments, involves substantial effort. Once a document collection is thus constructed, making a query set and annotating the dataset for relevance against each query demands time and labor. Thus a test collection is built. On this collection, initially, a language-identification-based solution with query expansion employing different phonetic algorithms is proposed. Subsequent analysis underscores the importance of devising domain-specific strategies, particularly in the context of social media, where the pattern of occurrence of stop words undergoes significant fluctuations over time. To this end, the development of a code-mixed stop words list is advocated to bolster the performance of information retrieval systems.

1.10 Structure of the thesis

The thesis consists of 8 different chapters. The structure of this thesis is delineated as follows.

Chapter 1 starts with a brief introduction to social media, code-mixing, task on CM data. The motivation, challenges, research goals and contributions of our work are also mentioned.

Chapter 2 provides an essential background information necessary for a comprehensive understanding of the thesis work is discussed.

Chapter 3 provides an extensive literature review, encompassing word language identification, sentiment analysis, hate speech identification, and information retrieval on code-mixed data.

Chapter 4 delves into the foundational aspect of any code-mixed task: word-level language identification.

Chapter 5 introduces a deep learning-based model tailored for sentiment analysis on Dravidian code-mixed data.

Chapter 6 showcases a deep learning-based model for identifying hate speech on Twitter data.

Chapter 7 focuses on elucidating the methodology employed in creating datasets for the code-mixed information retrieval task. First, the initial solution utilizing query expansion through phonetic encoding is discussed followed by description of an effective approach for generating stop words and consequential impact of stop words removal in CMIR.

Chapter 8 concludes encapsulating key findings and delineating open challenges and future directions for research in this domain.

A pictorial representation of the structure of this thesis organisation is illustrated in Figure 1.9.

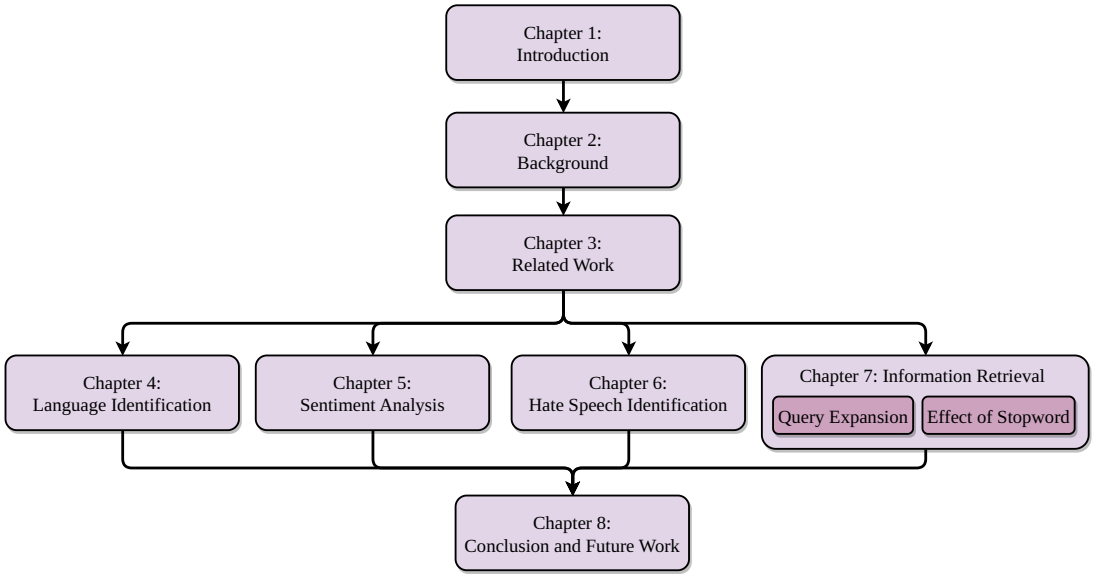


Figure 1.9: Overview of the dissertation's structure