University of Sheffield

Identifying complaints in social media using deep learning with transformers

Nitin Sunny Mathew

Supervisor: Nikolaos Aletras

A report submitted in fulfilment of the requirements
for the degree of MSc in Data Analytics

in the

Department of Computer Science

September 13, 2023

Declaration

All sentences or passages quoted in this report from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in this project and the degree examination as a whole.

Name: Nitin Sunny Mathew

Signature: Nitin Sunny Mathew

Date: 13-Sep-2023

Abstract

A complaint is a statement made by a person or an entity with the intent to indicate something is unacceptable or unsatisfactory. This is commonly used in various aspects of day-to-day life including when conducting business operations. With the proliferation of social media and their active enablement by organisations for user engagement, it has become a common medium for users to raise complaints. With such complaints being publicly visible, it becomes crucial for organisations to identify, prioritise and respond to these complaints swiftly. Automatically identifying complaints in social media is an active area of research. In the past few years, the focus has been on using NLP approaches driven by developments in transfer learning and transformer-based models.

This paper expands upon these methods by evaluating different variations of the BERT model. It includes BERTweet, which is pre-trained using tweets, as well as smaller models, DistilBERT and BERT Tiny which target to minimize the finetuning and inference time. Using a publicly available Twitter dataset, BERTweet achieves the highest performance with an F1 score of 0.908 while smaller models like DistilBERT and MobileBERT exhibit strong predictive performance for this task. Furthermore, the act of complaining and its nature when used online and in social media are analysed from a linguistic perspective along with discussions on state-of-the-art approaches for such NLP tasks.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In the act of complaining, dissatisfaction or annoyance is expressed by a person or entity in response to a previous or ongoing event that has negatively impacted them [27]. There is a breach of expectation and the act of complaining provides an avenue to direct dissatisfaction to the appropriate organisation or individual with the hope of rectification or redressal. It could also be used as a means to issue a 'Face Threatening Act' [6], to the detriment of the recipient's reputation. The event or action could be concerning a product or service procured by the concerned person or entity. The need to recognise, acknowledge and act on complaints is of significant importance to businesses and organisations to retain their customers while maintaining their reputations.

Until the advent of online platforms, specifically social media, the impact of negative online word-of-mouth was confined to a relatively limited audience. However, since then complaints posted online have the potential to rapidly go viral, reaching millions of individuals and significantly damaging a company's brand reputation and goodwill in a short period [39]. The multiple social media channels of organisations allow customers to express their complaints conveniently and with enhanced effectiveness, contributing to their increasing usage [2].

Examining instances of complaints on social media and specifically Twitter (rebranded as X[1]), they seem to be in alignment with the previously described act of complaining. These examples as shown in Table 1.1 are of individuals who have encountered breaches of their expectations. Regarding the intentions underlying these complaints, the objective is rectification in the first and second examples. In the first tweet, there is a request for a specific software version to resolve an issue, while the second tweet seeks clarification on a policy due to the perceived violation arising from a wrongly advertised product. In contrast, the third and fourth tweets are instances of issuing a 'Face Threatening Act'. They are written with the intention to harm the brand's value considering the use of terms such as *incompetence*, *worst*

[1]https://twitter.com/elonmusk/status/1683171310388535296

No.	Example complaints from Twitter
1	hi please i cant find a driver for video card (nvidia geforce 8500 gt) for mac please send me a link when i can download a driver
2	what is your policy on false advertising regarding sale items ? i was refused a sale in westfield due to a company error on pricing
3	thanks to <user> ' s incompetence i now can't work till october 4th , when the ati card arrives .
4	you jave the worst customer service #pissed #useless #worstbrand

Table 1.1: Sample complaints extracted from Twitter, exhibiting diverse degrees of complaint expression and severity. These complaints are sourced from data that has undergone the preprocessing steps outlined in Chapter 3.

customer service and hashtags like *#pissed*, *#useless* and *#worstbrand*. One could also argue that the second tweet encompasses elements of a Face Threatening Act, as mentioning false advertising in the context of a company could potentially harm the brand's reputation.

In addition to the timely addressing of customer complaints, automated detection of complaints in natural language has several other purposes. Linguists could gain an intricate understanding of the context, intent, and various types of complaints on a larger scale while psychologists could utilise this information to identify the underlying human traits that drive the behaviour and expression of complaints. Developing downstream natural language processing (NLP) applications, such as dialogue systems is another use case of this task [30].

Attempting to identify complaints manually through the multitude of posts and streams coming through the various social media channels is neither practical nor scalable. Various approaches to automate this task have been explored. The traditional vector-space method utilizing dictionaries has been applied in other text classification tasks [22]. Latent Semantic Indexing based on Singular Value Decomposition (SVD) along with linguistic style features to classify complaint emails was another approach [9]. In recent years, there has been a rise in the use of various Machine Learning (ML) and Natural Language Processing (NLP) based approaches for similar classification problems. The performance of logistic regression over various types of feature spaces against neural-network based models like Multi-layer Perceptron (MLP) and Long Short Term Memory (LSTM) has been analysed by [30]. The use of more advanced approaches using transformer networks has shown to have better results as explored by [18]. In this paper, the use of the BERT and its many variants, including that of recently created smaller variations will be assessed further on a publicly available Twitter dataset.

1.2 Objectives

- Evaluate the predictive performance of BERT and its variants on the complaint identification task and against the previous baseline.

- Analyse how the overall performance compares to the model sizes.
- Evaluate the behaviour of the best-performing model as characteristics of data used for finetuning vary. This includes changes in the volume, domain (industry type) and class balance.

Chapter 2

Literature Survey

2.1 The act of complaining

As per [27], the speech act of complaining in the traditional sense can be understood from the perspective of the speaker stating their displeasure or dissatisfaction to a target entity or individual. This is done as a reaction to an unfavourable event that is currently taking place or has already occurred. The authors believe a few preconditions have to be satisfied to result in a complaint being made. This includes the speaker's belief the entity or individual is responsible for the unfavourable outcome and that the speaker in question suffers from the consequences. The result is a verbally expressed complaint.

This expression of complaint could be carried out in various ways. The speaker might choose to directly communicate their complaints or concerns to the individual or entity, either immediately after the incident or at a later time. Or they might voice their grievances to others through word-of-mouth or they could even opt to escalate the issue by involving a third party, such as a consumer advocacy office [37].

The authors of [27] further delve into the intentions of the speaker in making the complaint. They argue this is carried out with either the hope of repairing the situation or as a 'Face Threatening Act' [6], with the purpose being to damage the face of the individual or entity against whom the complaint is made. In this scenario, a face-threatening act refers to an action that challenges the reputation of the recipient by going against what the recipient desires. These acts can manifest in a verbal form including with variation in tone or inflection or using non-verbal methods.

While such complaints could be considered direct complaints as per [5], the authors highlight the use of indirect complaining in speech. In the case of indirect complaints, the speaker does not attribute responsibility for the cause of the complaint to the individual or entity being addressed. The authors theorise, that an indirect complaint is used to bring about 'solidarity' between speakers, which is contrary to the use of direct complaints. It can serve as a means to

initiate conversations and establish temporary connections with others. The scope of the data for this project (described in the subsequent chapter) is primarily focused on direct complaints as they are selected based on tweets being addressed to a brand's customer service handle. However it is possible that tweets which fall into the category of indirect complaints are also part of the dataset.

Analysing deeper into which types of customers complain more, [36] have looked at how personality traits like impulsivity and self-monitoring impact customer complaining behaviour. *Impulsivity*, as defined by [33], refers to a consistent inclination of customers to act spontaneously and immediately, without much reflection or careful consideration of available options or potential consequences. This trait remains relatively stable over time for such customers. *Self-monitoring* is described by [3] as the propensity to adjust one's behaviour based on the actions or behaviour of others. High self-monitoring individuals are sensitive to others' expressions and behaviour, relying on social cues for their actions, while low self-monitoring individuals may be influenced by personal traits. From their experiments, [36] concluded that individuals with high impulsiveness tend to complain more than those with low impulsiveness, whereas individuals with high self-monitoring tend to complain less than those with low self-monitoring. These effects tend to be more pronounced in situations where the level of dissatisfaction is high. The level of *involvement* a customer has in a product or service also matters since the more deeply a customer engages with a consumption scenario, their inclination to invest resources like time, effort, and money into addressing or complaining about an unsatisfactory encounter increases [21]. The results from [36] also validated the positive influence a consumer's degree of involvement has on their likelihood of engaging in complaining behaviour.

2.2 Complaining online

The act of complaining exists online in various forms and with varying degrees of intensity and this prevalence led to the emergence of third-party organisations that provide online channels for customers' ease and convenience [39]. Notably, there are complaint websites like complaintsboard.com, review websites like trustpilot.com as well as consumer organisations' sites such as consumeraffairs.com, where customers can share their negative experiences and exchange information with others. The impact of negative word of mouth is quite high due to the ease with which negative reports can rapidly reach millions of people, potentially causing significant harm to a company's brand. Various user-generated content platforms such as YouTube, Twitter, and Facebook serve as spaces for expressing complaints. Brands use these platforms for user engagement and this provides the users with the required visibility to potentially raise or escalate an issue. With numerous such options available online, companies can experience significant repercussions arising from actions taken by dissatisfied customers [39].

Of the 431 online complaints assessed by [39], 96% followed what they call a double deviation. This occurs when customers experience both a product or service failure followed by multiple unsuccessful attempts to resolve the issue, resulting in them feeling they have been violated twice. Such customers then resort to online complaining. Their urge to complain online is driven by how they felt betrayed rather than simply being dissatisfied or with any form of malicious intentions to hinder business operations.

Complaining online is also associated with electronic word-of-mouth or EWOM, which involves sharing information online with a wider group, and it remains accessible over an extended period while often being anonymous [13]. This type of communication can take place on various platforms, ranging from official company-sponsored sites to unaffiliated blogs. Among the different forms of EWOM, consumer reviews are particularly noteworthy, as they provide valuable insights about products, whether positive or negative [37]. Such Negative electronic WOM (EWOM), can significantly damage a brand's reputation and influence potential customers to seek alternative products or services.

2.3 Complaining in social media

With the increased penetration of social media in the lives of consumers, they now have the ability to express their grievances directly and effectively to service providers using multiple social media platforms [2]. Prior to this, a significant portion of dissatisfied customers refrained from lodging complaints due to the perception that the costs associated with complaining far outweighed the benefits linked to resolution [36]. This has had a major impact at the fundamental level on how customers complain across industries. Aside from lodging complaints, consumers use social media channels to publicly vent their anger, a behaviour they formerly exhibited by privately expressing dissatisfaction to friends and family [2].

The study carried out by [2] investigated the elements that incite customers to participate in complaining activities via social media. They asserted that the act of complaining through social media was shaped by multiple factors. These include perceptions of not being treated fairly and wanting retribution, attributions of causality and the personal identity and traits of the individual. Particularly, they found that distinct factors impact both private and public complaints communicated through social media. They found the level of perceived unfairness in the situation had more impact on customers choosing to publicly complain through social media. Additionally, a positive view of the firm's compensation terms fostered the expectation that complaining increased the chance of resolution. Consequently, dissatisfied customers were more inclined to express their grievances privately to the service provider instead of publicly criticising them on social media.

Organisations have also been promoting the use of social media and digital channels as a medium for providing customer support while moving away from conventional contact points

like call centres [38]. The cost of handling per customer on Twitter, for example, is significantly lower when compared with a call centre. It is convenient for the customer as well since they can interact with a brand via the social media handles at their preferred time rather than having to wait for a prolonged time over a phone call. Effectively managing a customer's complaint also has the potential to transform a service failure into a favourable brand encounter. This brings about opportunities to showcase the brand in a positive light among other social media users.

The speed and scale of communication on social media is a key factor for opinion propagation online [29]. Posts on social media provide a continuous stream of information. When it comes to information that appeals to a wider audience, a considerable number of individuals can be reached swiftly. Consequently, this may lead to a situation where a particular topic gains prominence for a short while, triggering a surge in posts on that topic. The reality of social media demands swift responses from affected organisations, often within hours or minutes.

2.4 Twitter as a medium for self-expression

Twitter is a microblogging platform with over 540 million users and active users amounting to over 250 million[1]. It is known for being the fastest social media platform due to its high turnover of information and short message length [17]. This short and quick communication is influenced by both technical factors and the nature of interactions, where opinions are often simplified to likes or upvotes. Messages are usually limited in length enforced by Twitter's character limit (currently at 280 for most users[2]).

Different communication modalities exist for various social media platforms. Twitter uses short messages, primarily in text but could include images and videos while Instagram has images at its core, and Facebook combines both approaches along with privacy controls [35]. The study performed by [35] found that these varying aspects attract different types of consumers to these platforms. It also found that participants who were primarily Twitter users openly shared content leading to increased *Social Capital*. *Online Social Capital* while often quantifiable from the number of followers or connections and likes or views accumulates from digital interactions, which may or may not align with a user's perceived value in the online space[11]. Users utilise such indicators to achieve specific objectives as part of their online social life. Participants favouring Twitter appeared to exhibit stronger individualistic tendencies and gave importance to self-expression and choice. They also showed a greater inclination towards forming and dissolving relationships, and trust in unfamiliar individuals. They also found such users often set their profiles for public visibility while building up social

[1] `https://www.reuters.com/business/media-telecom/musk-says-x-monthly-users-reach-new-high-2023-07-28/`

[2] `https://developer.twitter.com/en/docs/counting-characters`

connections [35].

The authors of [32] explored how personality affects the content and tone of tweets. Microblogging on Twitter generates an extensive written record of daily behaviour in natural settings due to the active usage of the platform by millions of users. They found tweets to be containing valid linguistic cues to the user's personality. They suggest the linguistic characteristics of the tweets in terms of the length of the messages, variation in functional vocabulary and type of emotion used depends on the personality traits of the individuals including the level of extraversion, agreeableness and level of diligence.

2.5 Previous work on automation of complaints identification

The initial research, delving into a computational linguistic analysis of complaints in social media and their automated identification using ML was carried out by[30]. Beyond creating a publicly accessible and annotated Twitter dataset specifically for complaints, the authors conducted a broad evaluation of the quantitative features associated with complaint data. They investigated the efficacy of unigrams, Linguistic Inquiry and Word Count (LIWC) measures, and word clusters as potential features for the classification task. Furthermore, they explored how sentiment and emotion analysis could enhance the detection of complaints, given the intrinsic link between these factors and the language employed. Their findings demonstrated that models using logistic regression with all the extracted features yielded the best predictive performance, with an F1 score of 0.78. Interestingly, this performance surpassed that of neural network models such as Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM), which the authors contend could be due to the lower training data volume. Moreover, supplementary data collected via distant supervision was used to further enhance the Logistic Regression model's F1 score to 0.79.

This line of research was extended by [18] where they explored using pre-trained transformer language models[42], combined with linguistic information. They used BERT, ALBERT and RoBERTa for the assessment. M-BERT or Multimodal BERT was used to check how using additional linguistic information for emotion and topical themes could impact the predictive performance. They found the transformer models performed better than the previous baseline [30] and BERT performed the best with an F1 of 0.87 followed closely by RoBERTa. While M-BERT was competitive, the authors felt there was no benefit from injecting this additional information into the finetuning for the BERT models. They also found the distant supervision approach of [30] had a negative impact on the performance of the transformer models aside from M-BERT.

Complaints can be classified into different levels of severity using linguistic pragmatics. The severity depends on the extent to which the complainer's 'Face Threatening Act' is intended to harm the recipient. This categorization offers valuable insights into the motivations of the

customers and what organisations can do better in addressing the complaints [19]. The study by [19] explored the use of transformer models combined with linguistic information to perform a classification task for the severity level. The authors decided on the 4 severity levels of 'Disapproval', 'Accusation', 'Blame' and 'No explicit reproach' as defined by [40] and enhanced the data from [30] to annotate with the severity levels as the classes. Three baseline models were employed: the majority class, logistic regression with bag-of-words, and BiGRU-Att (a bidirectional GRU model with self-attention), alongside RoBERTa. Additionally, a modified version of M-BERT, termed M-RoBERTa was developed. Various versions of this model were then evaluated, incorporating linguistic data for emotions, topics, and a combination of both. The findings indicated that the M-RoBERTa models delivered the most promising results, suggesting that the inclusion of supplementary linguistic information indeed contributed to enhanced model performance. Following closely, the RoBERTa baseline showcased the next best performance, while the logistic regression and BiGRU-Att models faced more challenges in achieving competitive outcomes.

The automatic identification of complaints has been researched within other technology channels like email. In the study [9], the authors investigated methods to detect complaint emails in order to improve an organisation's complaint management capabilities. Their approach involved utilising a combination of document features represented as vectors (weighted term frequencies) as well as linguistic features. These features aimed to capture linguistic styles as predefined groups, and they also considered word counts within each group. To address the sparseness of the matrix, they applied dimensionality reduction using Latent Semantic Indexing. For classification, the authors employed boosting, a technique that combines the predictive outputs of multiple "weak" classifiers to create a more powerful ensemble of classifiers, expected to provide robust predictive performance [12]. The researchers built a corpus of emails received by a newspaper organisation's call centre. Based on their linguistic analysis, they discovered that complaint emails were more likely to exhibit characteristics such as using the present tense, containing higher word counts, articles, time-related indicators, and negations. In terms of predictive performance, their classification system demonstrated an AUC score ranging from 0.846 to 0.913 across different configurations, indicating its effectiveness.

2.6 NLP based solutions

Computational linguistics or Natural Language Processing (NLP) has evolved into a compelling scientific field with significant research being carried out while having practical usage across both consumer use cases as well business operations. NLP focuses on using computational methods to comprehend, generate, and learn from human language content [14]. The progress in this field over the last decade and more is attributed to four main factors by [14]: computing power advancements, access to extensive linguistic data, the success of advanced machine learning techniques, and a more comprehensive understanding of human languages in the

construct of social usage. The solutions target to solve problems in areas of machine translation, text generation, summarisation and classification, question answering as well generating business insight from the vast amounts of text available online [14].

2.6.1 Transformer network and inductive transfer learning

The authors of [42] introduced the 'self-attention' mechanism to build an auto-regressive transformer network made up of repeated encoder and decoder stacks but without recurrence as used by Long-Short Term Memory networks. Self-attention allows the network to identify and arrive at a representation of the sequence of text by looking at the entire input sequence at each step. This mechanism aids the model in developing a better understanding of the input data and the relationships between the various parts of the input.

[42] describe the encoder made up of a stack of identical layers and generate for an input sequence $[x_1, ..., x_n]$ a representation sequence $[z_1, ..., z_n]$. This representation form is utilised by the decoder which is also stacked to generate an output sequence as $[y_1, ..., y_m]$. At each stage, the decoder uses the previously generated output sequence to create the next token in the sequence making the model auto-regressive. The authors also introduced the concept of multi-head attention where instead of performing attention just once at each step it is performed multiple times (8 were used in the paper). However, the dimensionality is reduced such that when the final output of all the attention projections is concatenated they have the same dimensionality as single-head attention. This limits any impact on computation cost. The results from their experiments showed this architecture had a performance advantage over previous BLEU[3] benchmark scores for English-to-German and English-to-French translation tasks. They also showed significant savings in terms of training computation cost due to the level of parallelisation that was achieved [42].

Inductive transfer is a transfer learning technique in machine learning to leverage the training process and apply learnings and representations from one task onto a different task that has a different data distribution [16]. Considering the limited success of transfer learning in NLP tasks, [16] identified key challenges, reflecting on past research. While computer vision tasks had found success in implementing inductive transfer learning, the authors felt some of the issues faced by NLP tasks were induced by researchers not appreciating the differences between the two task categories. Recognizing potential benefits in tailoring approaches, they proposed a method that pre-trained a language model on Wikitext-103 [25] representing the general domain dataset. This was followed by fine-tuning the language model using data from the various domains scoped for their experiments. For their network, they decided on 2 layers for task-specific classification which were fine-tuned using data from the target task's domain. Their approach was able to achieve state-of-the-art results on tasks for sentiment analysis,

[3]Bilingual Evaluation Understudy, an evaluation method for machine translation of human language.

question answering and topic classification.

2.6.2 BERT and its variants

Bidirectional Encoder Representations from Transformers or BERT was proposed by [10] to solve some of the issues with the transformer networks due to their use of unidirectional language models during the pre-training phase which limits the strengths of pre-trained representations. They proposed using two pre-training tasks. The first was a masked language model or MLM where a token has been randomly masked and the objective is to predict based on the context. They argued this would assist in pre-training deeper and bi-directional transformer models. The second task involved pre-training sequence pairs and was termed as next sentence prediction or NSP. They showed that using very large model sizes could also result in notable enhancements for even smaller downstream tasks, as long as the model was adequately pre-trained. Their experiments were run with 2 versions of BERT, BERT Base with 110M parameters and BERT Large with 340M parameters. The BERT models demonstrated superior performance compared to other state-of-the-art models across various benchmarks encompassing tasks related to natural language comprehension, inference, and question-answering.

With the significant success of the BERT Transformer model, several derived models were proposed over the next few years with some of them focusing on improving the pre-training while others were on reducing the model size. Some of the important ones are covered in brief. RoBERTa or Robustly Optimised BERT Pretraining Approach was introduced by [23] to overcome what the authors felt were shortcomings in BERT's [10] pretraining strategy including certain design choices. They made changes in terms of learning rate warmup and tuning the Adam optimiser. They also increased the pretraining with an additional corpus. With these changes, the model achieved state-of-the-art benchmark results.

While larger models with appropriate pretraining have better predictive performance [10, 23] there exist challenges posed by memory limitations and training times associated with model size increases. A Lite BERT or ALBERT was introduced by [20] to include techniques to enhance the efficiency of large-scale pre-trained language models. ALBERT employed two methods: factorised embedding parameterisation and cross-layer parameter sharing. The former divided the embedding matrix into smaller components, allowing hidden layers to expand without a substantial rise in vocabulary embedding parameters. The latter controlled parameter growth as network depth increased. These approaches significantly reduced parameters while maintaining performance. ALBERT with BERT Large like configuration has 18 times fewer parameters and 1.7 times faster training time. Their experiments showed ALBERT xxlarge, with about 70% of BERT Large's parameters had significant improvements compared to BERT-large across benchmark downstream tasks with speedup achieved in training times as well.

DistilBERT [34], a variation of BERT with the knowledge distillation technique applied in the pretraining phase was another model proposed to address similar challenges. Knowledge distillation involves distilling the learnings of a larger 'teacher' model (or ensemble) to a smaller and compact 'student' model as part of training and hence is a form of a compression technique [34]. The results from the experiments showed this model had 40% fewer parameters than BERT Base but was able to achieve performance on downstream benchmark tasks within just 0.6% to 3.9% points behind BERT Base. Another goal of the authors was a reduction in inference time, which they achieved by the model being faster by up to 60%.

To develop models that align to pre-defined memory and computational speed requirements, [41] proposed various versions of BERT with changes in the model size and embeddings size. They used various strategies to train a student compact model, combined with a bigger teacher model. The smallest model was Transformer Tiny (BERT Tiny) with 4.4M parameters while the largest was BERT Base with 110M parameters. They used various techniques to develop the compact model and compared them with BERT Large as the baseline on 3 NLP tasks. Pretraining the compact model with a masked language model objective and following it with knowledge distillation provided the best performance. This approach was superior to using distillation alone or relying on pretraining combined with fine-tuning.

To conclude this section, an approach of pre-training BERT on English tweets to enhance performance on Twitter data is examined. BERTweet Base was introduced by [26] as the first pre-trained model on large-scale tweets dataset in the public domain. While retaining the BERT Base architecture they used the pre-training optimisations introduced by RoBERTa [23]. They argued the linguistic attributes of tweets like length, grammar and vocabulary differ from the text present in the corpora used for pre-training BERT. The pretraining was performed using a corpus of 850M tweets. This was followed by fine-tuning the model independently for each of the 3 downstream Twitter NLP tasks. Their results showed the model did better than RoBERTa Base in all 3 NLP tasks. While RoBERTa Large did outperform BERTweet Base on the POS tagging task, they speculate this could be attributed to the significant difference in model sizes.

2.6.3 Ongoing research

GPT or Generative Pre-trained Transformer models have significantly changed the landscape in terms of the general perception of AI as well as its applicability. As per OpenAI, the developers of ChatGPT and other GPT-[2,3,3.5,4] versions, of their models have undergone training to comprehend both natural language and code. These models generate textual responses based on their inputs. The inputs given to GPT models are commonly referred to as 'prompts'. GPTs enable the development of applications for tasks such as writing documents, coding, answering questions, text analysis, creating dialogue agents and language

translation among others[4]. The use of prompts allows for performing text classification tasks as well. Some of the recent research on such models will be discussed next.

When using transformers, one of the primary drawbacks is that despite the task-agnostic nature, datasets relevant to that task along with fine-tuning are still essential for good performance [7]. This dependency on extensive labelled datasets limits the wider application of language models. The authors contend the aspiration for NLP systems is to eventually achieve versatility and broadness similar to that of humans who typically require minimally supervised datasets to grasp most language tasks. Meta-learning, where a language model learns a wide skill set during training and applies them at inference to adapt quickly to tasks is an approach researched in recent years [7]. Despite some initial promise, this approach lags fine-tuning in terms of results.

Another promising direction involves increasing the size of the transformer language models, leading to enhanced NLP performance [7]. The authors argue due to the nature of meta-learning, larger models might exhibit significant improvements in in-context learning abilities. To validate this, the researchers built GPT-3 [7], an auto-regressive model with 175 billion parameters. GPT-3 was evaluated across three conditions, 'few-shot learning' with multiple demonstrations within the context window, 'one-shot learning' with a single demonstration, and 'zero-shot' learning, relying solely on natural language directions. The model demonstrated good performance across NLP tasks and benchmarks with the 3 evaluation conditions. It closely rivalled state-of-the-art fine-tuned systems in certain cases. For ad-hoc tasks, the model exhibited high qualitative results. Finally, the authors assessed the social impact of using such models and highlighted the models show varying levels of bias across race, religion and gender. They feel this is likely arising from the inherent stereotypes captured within the training data sourced from the internet [7].

Capable of processing both images and text as input, GPT-4 [28] is a large transformer-based multimodal model, that generates text as its output. It is pre-trained to predict the next token as the output. Such models have a broad area of application with machine translation, text summarisation and dialogue systems as some of the main use cases. The authors [28] claim the strength of GPT-4 is in understanding and generating natural language in intricate and complicated situations and show it performs highly in exams designed for humans. While the model is superior in performance to other state-of-the-art systems in both English and other languages it suffers from limitations including 'hallucinations', inability to learn from experience, restricted context and inherent bias. While GPT-4 alleviates the hallucination problem when compared to GPT-3.5 (successor to GPT-3), authors suggest usage with caution, especially in mission-critical scenarios. GPT-4 can handle prompts containing both images and text, allowing users to define various vision or language tasks. The model generates text outputs based on mixed text and image inputs, displaying comparable proficiency across

[4] `https://platform.openai.com/docs/guides/gpt`

different domains as it does with text-only inputs [28].

While the model is extremely versatile and outperforms other state-of-the-art models in tests, the authors acknowledge the risks associated with the higher capabilities and are working with other researchers on recommendations on how society can prepare itself for any potential social and economic impacts from the wider usage of such AI technologies [28].

Chapter 3

Methodology

3.1 Task

For a short text segment, $T = \{t_1, t_2, ..., t_n\}$ where t_i is defined as a token, classify if the sequence of tokens T is a complaint or not.

3.2 Data and pre-processing

The data used for the experiments is from Twitter which provides a good representation of social media text due to the direct connection consumers have with organisations and brands [30]. With Twitter, users have a medium where they can self-express to a higher degree while maintaining flexible connections across their social network[35]. The data set created by [30] and further used by [18] is utilised for this project. The original process for collection and annotation employed by them is briefly described below. The particular version[1] used for the experiments is the one enhanced by [19] with the addition of labels for the severity of complaints. These additional labels are not used for the experiments in this project.

3.2.1 Criteria for tweets

A cross-industry representative collection of 93 customer service handles of organisations on Twitter were identified manually. These handles were then categorised into 9 domains using their industry type. Since an organisation could have business activities across domains, the assigned domain was based on the products or services receiving the most number of complaints. All the domains used in the experiments are listed in Table 3.1.

3.2.2 Data Extraction

Twitter API[2] was utilised to extract the tweets. The latest 3,200 tweets at the time of the collection exercise were retrieved and the original tweets to which the customer service handles

[1] `https://archive.org/details/complaint_severity_data`

[2] `https://developer.twitter.com/en`

Domains	Complaints	Non-Complaints	Total Tweets
Food & Beverage	95 (73%)	35 (27%)	130 (7%)
Apparel	141 (55%)	117 (45%)	258 (13%)
Retail	124 (62%)	75 (38%)	199 (10%)
Cars	67 (73%)	25 (27%)	92 (4%)
Services	207 (61%)	130 (39%)	337 (17%)
Software & Online Services	189 (65%)	103 (35%)	292 (15%)
Transport	139 (56%)	109 (44%)	248 (12%)
Electronics	174 (61%)	112 (39%)	286 (15%)
Other	96 (79%)	33 (21%)	129 (7%)
Total	1232 (63%)	739 (37%)	1971

Table 3.1: The nine domains and the distribution of tweets that are complaints and those that are not. The percentages indicate how the splits are distributed [30].

responded were identified. Random sampling equally for each handle, 1,971 tweets were then identified where there was a response from the support's handle. To ensure a more balanced and diverse dataset, 1,478 randomly sampled tweets were added to the dataset. 739 tweets were replies to other handles (outside the 93 identified) and the remaining 739 tweets were not addressed to any Twitter handle. Table 3.2 shows the breakdown of the total population of the tweets dataset. Tweets were filtered for English using langid.py [24]. Retweets were excluded and all usernames and URLs were anonymised and replaced with placeholder tokens.

Extraction Criteria	Complaints	Non-Complaints	Total Tweets
Addressed to and replied by the identified 93 customer service handles	1239 (63%)	739 (37%)	1971 (58%)
Addressed to other customer service handles	0	739 (100%)	739 (21%)
Not addressed to any Twitter handle	0	739 (100%)	739 (21%)
Total	1232 (36%)	2217 (64%)	3449

Table 3.2: Selection of tweets based on random sampling and where they have received replies when addressed to the 93 customer service handles combined with random sampled tweets that are addressed to other handles (random_reply) and tweets that are not addressed to any handle (random_tweet) [30].

3.2.3 Annotation

The classification of the 1,971 tweets as complaints or not was carried out using a binary annotation task (complaint or not). Since tweets are concise and typically express a single idea, an entire tweet was classified as a complaint if it contained at least one speech act of complaining. To guide the annotation process, a complaint definition from [27], stating that a complaint portrays a situation that contradicts the writer's positive expectation was used. Two of the authors with extensive annotation experience in linguistics independently labelled the 1,971 tweets. They had substantial agreement [1] with Cohen's Kappa of $\kappa =$

0.731. In the end, 1,232 tweets (63%) and 739 tweets (37%) were identified as complaints and non-complaints. Table 3.1 gives the breakdown of the complaint and non-complaint tweets for each domain.

3.3 Environment

The key details of the environment used for the experiments are listed below. All experiments are run in a Jupyter Notebook and on a single GPU.

3.3.1 Hardware

- **CPU Count:** 8
- **System Memory:** 45 GB
- **GPU Count:** 1
- **GPU Model:** NVIDIA RTX A4000[3]
- **GPU Memory:** 16 GB

3.3.2 Software

For the experiments, the BERT large language model along with a number of its variants are used to classify the tweets and compare the performance. The models are based on the `transformers` library implementation from Hugging Face[4] [43] in Python. Additionally, the `datasets` and `evaluate` libraries are used. From scikit-learn[5] the `sklearn` library is utilised to generate the stratified splits for the nested cross-fold validation. The versions for each library are shown in the table 3.3. Python version used is v3.9.16.

Provider	Library Name	Version
Hugging Face	transformers	4.21.3
	datasets	2.4.0
	evaluate	0.4.0
Scikit-Learn	sklearn	1.1.2
Numpy	numpy	1.23.4
Pandas	pandas	1.5.0

Table 3.3: Software and library versions used for this project. Other more

[3] `https://www.nvidia.com/en-gb/design-visualization/rtx-a4000/`
[4] `https://huggingface.co/`
[5] `https://scikit-learn.org/stable/`

3.4 Model selection

The performance of BERT and its variants on the text classification task will be explored as part of the experiments. BERT [10] is based on the modern transformers network architecture [42]. Using BERT for the text classification task has several advantages over previous dominant methods such as Gated Recurrent Units GRU) [8] or Long Short Term Memory (LSTM) [15] networks. Although tweets tend to be made up of short texts, the ability to capture long-term dependencies is still useful for understanding relationships across the content better. They also rely on bidirectional processing to use contextual information to have a more nuanced understanding of the intention of the author of the tweet or post. Since BERT is already pre-trained on large corpora, it is considered to possess a significant general understanding of language. Finally, the pre-training enables transfer learning and domain adaptation with relative ease which is very useful for tasks where annotated data is limited (only 3,449 tweets are used for the experiments).

The transformer models used are listed in table 3.4 along with the number of parameters for each of them. The number of parameters or model size is based on the embedding and output layers along with the attention heads. The models chosen are such that there is a wide range of model sizes, from RoBERTa Base and BERT Base with 125 and 110 million parameters to lightweight variants such as DistilBERT Base and BERT Tiny with much lower model sizes. This allows for a comparison of the model performance both in terms of the predictions as well as the inference time to the model size. Various studies have pointed to a relation between the size of the model and its predictive performance conditional to appropriate pre-training [42, 23]. The summary of BERT Base and BERT Tiny are shown in Figure 3.1 for a high-level perspective on the layers and the number of parameters in each of them.

Model	**Parameter Count**	**Tokenizer Type**	**Vocab. Size**
RoBERTa Base	125M	Byte-level BPE	50,265
BERT Base (uncased)	110M	WordPiece	30,522
BERTweet Base	110M	Byte-Pair Encoding (BPE)	64,000
DistilBERT Base (uncased)	66M	WordPiece	30,522
ALBERT Base v2	11M	SentencePiece	30,000
BERT Tiny	4.4M	WordPiece	30,522

Table 3.4: The transformer models used for the experiments along with the type of tokenization, and vocabulary size and sorted by the number of parameters for each of them. The parameter counts are from [4] for RoBERTa Base, BERT Base, ALBERT Base and BERT Tiny. For BERTweet it is from [26], and DistilBERT from [34]. Vocabulary sizes and tokenizer types are based on documentation at `https://huggingface.co/`.

```
==========================================================  ==========================================================
Layer (type:depth-idx)                   Param #            Layer (type:depth-idx)                   Param #
==========================================================  ==========================================================
├─BertModel: 1-1                        --                 ├─BertModel: 1-1                        --
│    └─BertEmbeddings: 2-1              --                 │    └─BertEmbeddings: 2-1              --
│    │    └─Embedding: 3-1              23,440,896         │    │    └─Embedding: 3-1              3,906,816
│    │    └─Embedding: 3-2              393,216            │    │    └─Embedding: 3-2              65,536
│    │    └─Embedding: 3-3              1,536              │    │    └─Embedding: 3-3              256
│    │    └─LayerNorm: 3-4              1,536              │    │    └─LayerNorm: 3-4              256
│    │    └─Dropout: 3-5                --                 │    │    └─Dropout: 3-5                --
│    └─BertEncoder: 2-2                 --                 │    └─BertEncoder: 2-2                 --
│    │    └─ModuleList: 3-6             85,054,464         │    │    └─ModuleList: 3-6             396,544
│    └─BertPooler: 2-3                  --                 │    └─BertPooler: 2-3                  --
│    │    └─Linear: 3-7                 590,592            │    │    └─Linear: 3-7                 16,512
│    │    └─Tanh: 3-8                   --                 │    │    └─Tanh: 3-8                   --
├─Dropout: 1-2                          --                 ├─Dropout: 1-2                          --
├─Linear: 1-3                           1,538              ├─Linear: 1-3                           258
==========================================================  ==========================================================
Total params: 109,483,778                                  Total params: 4,386,178
Trainable params: 109,483,778                              Trainable params: 4,386,178
Non-trainable params: 0                                    Non-trainable params: 0
==========================================================  ==========================================================
```

Figure 3.1: A summary of the BERT Base (left) and BERT Tiny (right) models for comparison in terms of the layers used and number of parameters. This is based on the models from Hugging Face.

3.5 Data tokenisation

The tokenization process is required for appropriate preparation of input data for use by BERT and its variants. The tokenization process involves dividing the input text into tokens based on a predefined set of rules. These tokens are subsequently transformed into numerical representations and tensors, along with any extra inputs needed by the model. Tokens in general could be words, subwords, phrases or even characters. There are various approaches to tokenization and the methods used by each of the models in the scope of the experiments are indicated in Table 3.4. The `transformers` library provides relevant methods to tokenize the input tweets. The library includes model-specific tokenizers such as, `BertTokenizer` [6] or `RobertaTokenizer` [7] while models like BERT-tiny leverage existing ones. For the experiments, the `AutoTokenizer` [8] has been used which conveniently selects the appropriate tokenizer relevant for the model in use.

3.5.1 Choice of settings

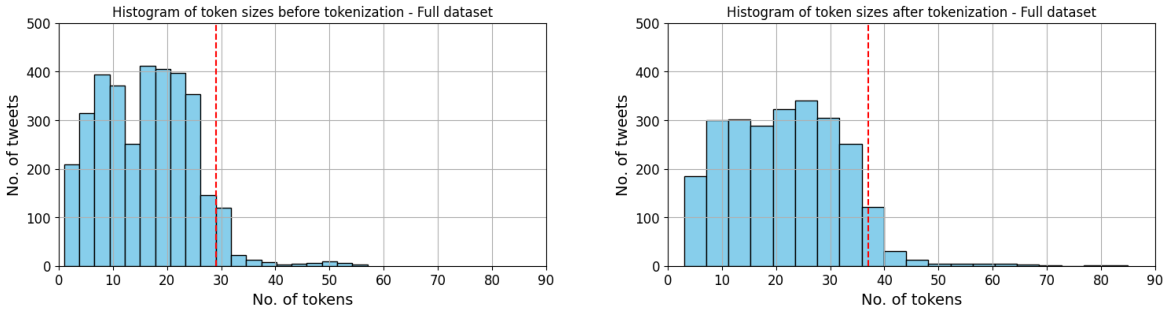
Prior to applying tokenization, the settings for padding and truncation [9] are chosen to ensure the varying input length will still result in rectangular tensors. The parameter, `max_length` determines the maximum number of tokens for each input while `padding` controls the type of padding and `truncation` allows to truncate input to a pre-determined number of tokens.

[6] `https://huggingface.co/docs/transformers/v4.21.3/en/model_doc/bert#transformers.BertTokenizer`

[7] `https://huggingface.co/docs/transformers/v4.21.3/en/model_doc/roberta#transformers.RobertaTokenizer`

[8] `https://huggingface.co/docs/transformers/v4.21.3/en/model_doc/auto#transformers.AutoTokenizer`

[9] `https://huggingface.co/docs/transformers/pad_truncation`



(a) Before tokenization

(b) After tokenization

Figure 3.2: The token count distribution for the full dataset of 3,449 tweets before and after tokenization with the red dashed line indicating 95% coverage of tweets. BertTokenizer is used here.

The distribution of the number of tokens in the tweets from the pre-processed data from [30, 19] before applying the model-specific tokenization is shown in Figure 3.2a. Over 95% of the tweets have 29 tokens or less. Using the BertTokenizer as an example, from Figure 3.2b it was found about 37 tokens are required to comprehensively cover 95% of the tweets. The other tokenizers require between 35 and 43 tokens to cover the same percentage (refer Appendix A.3). This analysis assists in the decision on the appropriate `max_value` for the tokenizer. A value of 50 ensures coverage of over 99% of the tweets completely for all the tokenizers. This when used in conjunction with `truncation=True`, sets the maximum number of tokens for each input tweet to 50. Anything that follows is truncated and not used for training or inference. Additionally, since the dataset includes shorter tweets with resulting tokens less than 50, `padding` is set to *'max_length'* to apply padding up to 50 tokens.

3.5.2 Tokenisation example

An example tweet from the input data is shown in **A**. The pre-processing applied by [30, 19] results in punctuation as separate tokens, e.g. 'again' and '.'. The hashtags are retained as single tokens. After applying tokenisation using the `BertTokenizer`, the data is converted into a list of input IDs representing their reference into the model's vocabulary as shown in **B**. To better understand the effect of tokenization, **C** shows the decoded input from **B**. The tokenizer adds special tokens, `[CLS]` - classification token for the beginning of an input sequence, `[SEP]` - separator token to separate input sequences and `[PAD]` - padding token. Aside from this, punctuation is combined with the word for 'again.'. In the case of hashtags, the '#' symbol has been separated out as a token.

A - Tweet from input dataset

```
love it when i almost die rear ended by a semi cause my jeep turns off again
.  one day they will fix it #jeepsucks #chrysler
```

B - Encoding the input

```
[101, 2293, 2009, 2043, 1045, 2471, 3280, 4373, 3092, 2011, 1037, 4100, 3426,
2026, 14007, 4332, 2125, 2153, 1012, 2028, 2154, 2027, 2097, 8081, 2009,
1001, 14007, 6342, 10603, 1001, 17714, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0]
```

C - Decoding the tokenized input

```
[CLS] love it when i almost die rear ended by a semi cause my jeep turns off
again.  one day they will fix it # jeepsucks # chrysler [SEP] [PAD] [PAD]
[PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
[PAD] [PAD] [PAD]
```

3.6 Experiments set 1: Predictive performance comparison of BERT variants

In the first experiment, the objective is to identify which of the models performs the best for the text classification task of complaints identification. Additionally, the relative performance of the models and the inference time will be analysed to assess how the model size impacts these aspects. A nested cross-validation approach will be used to experiment finetuning the model the various learning rate hyperparameter values and calculate mean performance metrics on inference. All models in scope for the experiments are pre-trained, hence finetuning will be performed for the downstream complaints identification task.

The nested cross-validation approach utilized is adapted from [30]. The outer loop consists of 6 iterations and the inner loop of 4 iterations. Each outer loop includes a stratified split of the entire dataset into train (**A**) and test (**B**) datasets. Within the inner loop, **A** is further split into inner train and dev datasets using stratified split with each iteration, finetuning and validating on 1 of the 4 learning rates set for hyperparameter tuning. The best model from the inner loop is selected based on the F1 score on the dev dataset. This best model is used to perform inference using the test dataset, **B**. At the end of the 6 outer loop iterations, the mean of the performance metrics is calculated as the final metrics for that model. While [30] used 10 iterations for the outer loop, it has been restricted to 6 for the experiments considering significant variations were not observed on the metrics. This is likely due to the smaller size of the dataset and 6 stratified splits capturing sufficient variability in the input dataset.

The key choices for the experiments including that of the hyperparameters are described in Table 3.5. For the stratified split, the `StratifiedKFold` function from `sklearn` library is

Parameter	Value
Outer loop iterations	6
Inner loop iterations	4
Random Seed	2023
Hyperparameter	**Value**
No. of Epochs	4
Learning Rate	[1e-5, 5e-6, 5e-5, 3e-5]
All other hyperparameters	Model defaults

Table 3.5: The choice of key parameters and hyperparameters used for Experiment 1.

used. This results in approximately 2874 and 575 tweets for the outer loop's train and test datasets and 2155 and 719 tweets for the inner loop's train and dev datasets. The number of epochs is set to 4 in line with official documentation for BERT [10] where they use between 2 and 4 epochs for the various downstream tasks. The learning rate includes the default rate used by the models as well as a range of alternate values. All other hyperparameters take on their default values for the models as defined in the `transformers` library.

For the predictive performance metrics, precision, recall, accuracy, F1 and Area under the ROC Curve (AOC) scores are computed for both the inner loop's validation as well as the outer loop's testing. Further, the final metrics for each model are based on each of the mean metrics from the 6 outer loop iterations. Additionally, the samples per second and steps per second are captured for each of the models during the inference phase to analyse the time taken for inference in relation to the model size.

3.7 Experiments set 2: Cross-domain predictive performance comparison

Cross-domain experiments are conducted to understand how the performance of finetuned models varies as the volume, domain and class balance of the tweets change. This approach also offers an avenue to assess any linguistic variations among complaints across different domains and their consequent effects on the classification performance.

For the experiments, the best-performing model from the first experiment is utilised. The tweets for each of the 9 domains are used for training separately with tweets from every other domain used for testing. Additionally, each domain is used for testing while training is based on all tweets except the domain used for testing. Stratified split is applied on the training dataset for each domain using the `StratifiedKFold` function from `sklearn` library for 3 iterations of finetuning. At the end of each iteration, inference is performed on the testing data. The number of epochs remains at '4', similar to experiment 1 while the learning rate used is the best learning for the selected model from experiment 1. All other hyperparameters

[10] `https://github.com/google-research/bert`

are the model defaults. The parameters and hyperparameters used for the experiment are listed in Table 3.6.

In this set of experiments, only the most effective model identified from the previous experiment is utilised. The model is finetuned using tweets belonging to one domain at a time while testing is performed for each of the remaining domains separately and performance recorded. Additionally, each domain is tested on a model which is finetuned on all tweets except the domain used for testing.

To ensure the balance of classes is maintained for the training and evaluation sets, a stratified split is applied using the `StratifiedKFold` function from the `sklearn` library. This process is repeated for three iterations of finetuning. Following each iteration, the model's performance is evaluated through inference on the testing data. The experiment maintains a consistent number of epochs of 4, similar to the first experiment. The learning rate is determined by the optimal value identified in the initial experiment for the selected model. All other hyperparameters follow the model's default settings. For details of the parameters and hyperparameters employed in this experiment, please refer to Table 3.6.

Parameter	**Value**
No. of iterations	3
Random Seed	2023
Hyperparameter	**Value**
No. of Epochs	4
Learning Rate	Best learning rate from Experiment 1
All other hyperparameters	Model defaults

Table 3.6: The choice of key parameters and hyperparameters used for Experiment 2. Refer to the Chapter on results for the model and learning rate used for this experiment.

For each iteration and combination of domains for finetuning and testing, the prediction performance metrics are calculated for precision, recall, accuracy, F1, and AOC. At the end of the third iteration, the mean of the metrics is calculated. The results and findings from this set of experiments are presented in Chapter 4.

3.8 Ethical, Professional and Legal Issues

The data used for the experiments were created by [30] and further enhanced with complaints severity type annotation by [19]. This data is anonymised and is available in the public domain[11]. No additional data has been collected for the experiments conducted for this project. To ensure the appropriate compliance with the ethical review requirements of the University of Sheffield, a self-declared ethics review application with reference number 054854 was raised and subsequently approved by the University Research Ethics Committee.

[11] `https://archive.org/details/complaint_severity_data`

Chapter 4

Results and discussion

4.1 Data exploration

As described in Chapter 3, the data for the experiments is taken from Twitter. It was extracted and pre-processed by [30] and further enhanced with the labels for complaint severity by [19]. What follows are the key findings from the exploratory data analysis performed. Some minor differences in the distribution of the tweets across the domains are observed between the latest version of the dataset available in the public domain[1] and the distribution described in the original paper. Since the variations are minor (0.5 to 2%), any potential impact on the model performance should be insignificant in the context of the objectives of the experiments. Refer A.1 for the full breakdown of the dataset used here.

4.1.1 Domain and class distribution

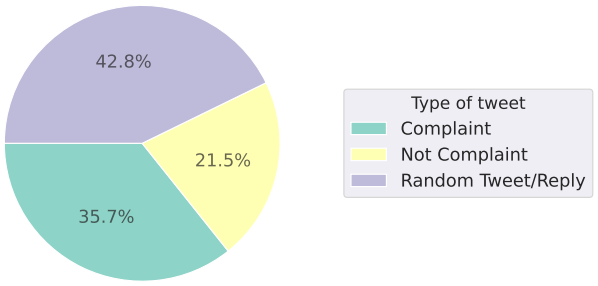
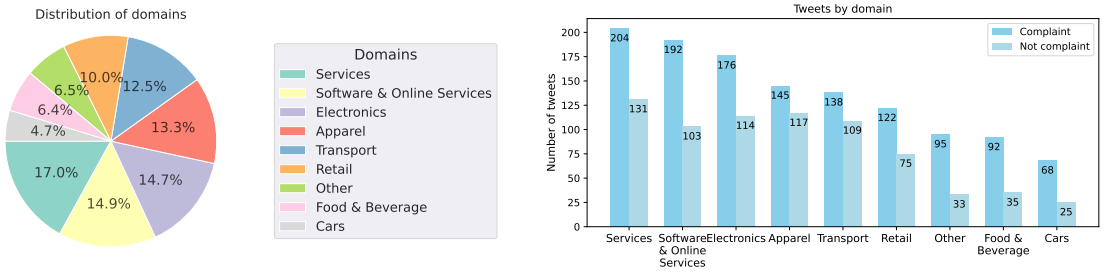


Figure 4.1: Illustrates the distribution of tweets categorised as 'complaints' and 'not complaints', with random 'tweets / replies' shown separately.

All tweets categorised as complaints are assigned `label:1`, while tweets that do not constitute complaints are assigned `label:0`. In terms of class distribution, the dataset is skewed towards 'not complaint' tweets, as depicted in Figure 4.1, where `label:1` represents 35.7% and `label:0`

[1] `https://archive.org/details/complaint_severity_data`

represents 64.3% of the dataset. Random tweets and replies with `label:0` were added by the authors of [30] to ensure a more representative dataset. This approach aligns with the real-world scenario where complaint-related posts form a smaller proportion within an organization's social media tweets and posts. Additionally, this strategy has the potential to enhance the model's ability to generalize effectively during the finetuning process.

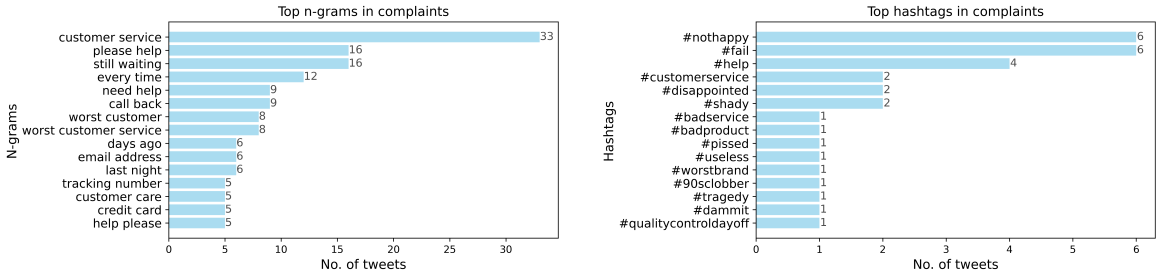


(a) Proportion of each domain

(b) No. of tweets for each domain

Figure 4.2: Shows the distribution of the domains used in the dataset

The dataset comprises domains encompassing both complaint-related tweets and non-complaint tweets. Figure 4.2a illustrates the distribution of domains, with the top 3 categories being services, software, and electronics, collectively constituting nearly 50% of the tweets. A key observation from Figure 4.2b is the prevalent class imbalance within most domains, accompanied by relatively low tweet volumes within each domain. The implications of these observations on predictions are analyzed in Experiments set 2 and elaborated upon later in this chapter.



(a) Top n-grams

(b) Top hashtags

Figure 4.3: Top phrases as n-grams(excl. unigrams) and hashtags in the complaint tweets.

4.1.2 Linguistic analysis

Delving deeper into the language used in Twitter complaints, the top phrases are analysed by extracting n-grams. As depicted in Figure 4.3a, common phrases in the complaint tweets either convey an exepectation for resolution (such as "please help" and "need help") or express

frustration (like "still waiting," "worst customer service," and "call back"). Others cover broader customer service themes (for instance, "tracking number" and "customer care"). To elaborate further, sample tweets with these phrases are shown below. They showcase various characteristics previously discussed, including instances of Face Threatening Acts, feelings of betrayal, altruistic behaviour (warning others), as well as elements like sarcasm. These findings align with the definition of a complaint and the intentions of the speaker as outlined in previous chapters.

Examples for expectation of rectification

> *"hey chrysler cares i'm the one with the 2011 200* ***need help*** *with the heating . inside the car it's really strange"*
>
> *"can someone* ***please help*** *me ? i've already sent a dm ."*

Examples for expression of frustration

> *"****worst customer service*** *experience with <user> <user> <user> . never been treated with such contempt"*
>
> *"on hold with <user> an hour just to get told to* ***call back*** *another day . hell yeah"*
>
> *"****worst customer service*** *to-date <user> in greensboro off wendover . avoid this place and let's show them we have other choices . #otherchoices"*

Examination of the hashtags within the complaint tweets as shown in Figure 4.3b points to their usage predominantly as a means of conveying frustration. Hashtags such as `#nothappy`, `#fail`, and `#disappointed` are examples. Consequently, in addition to expressing dissatisfaction, these hashtags also communicate negative sentiments. Apart from these particular types of hashtags, various brand-specific or product-specific hashtags are used. As per Twitter[2], users utilize the symbol "#" (hashtag) preceding a keyword or phrase significant to the context in their tweet to classify those tweets, facilitating their visibility in Twitter searches. Clicking or tapping on a hashtagged term within any message reveals additional tweets containing the same hashtag. Hashtags can be inserted at any point within a tweet. Frequently, words marked with hashtags that attain significant popularity transform into trending topics.

However, the volume of tweets which include hashtags, particularly in the context of complaints is relatively low. Out of a total of 459 tweets, only 149 complaint tweets incorporate hashtags, constituting roughly 12% of all complaint-related tweets. When excluding random tweets and replies, the tweets that are not complaints containing hashtags amount to only 67 instances. There is an average of 1.56 hashtags in tweets containing at least one. While the inclusion of

[2] `https://help.twitter.com/en/using-twitter/how-to-use-hashtags#`

hashtags may offer some assistance to the predictions by the models, their overall impact on the fine-tuning process could be quite limited due to their low prevalence in the dataset.

4.1.3 Sentiment analysis

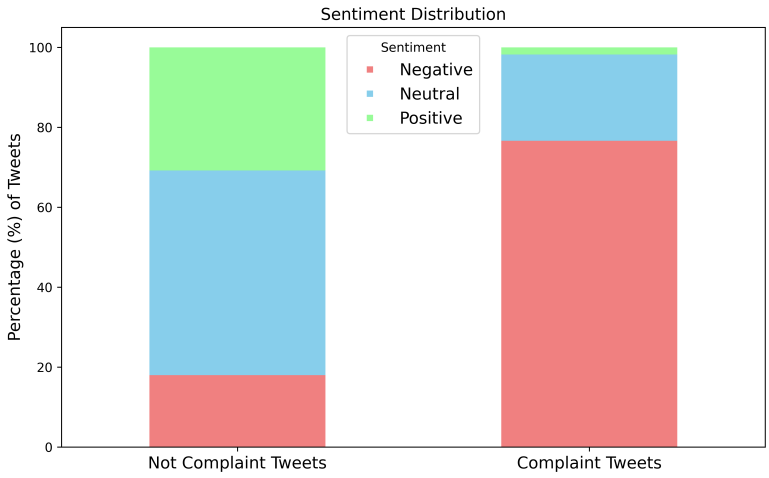


Figure 4.4: Distribution of positive, negative and neutral sentiments in the tweets.

The act of complaining typically involves conveying a sense of negativity. Sentiment analysis was conducted on the dataset using pysentimiento's library [31] and the results are in Figure 4.4. As expected, the majority of complaint-related tweets convey some form of negative sentiment, while approximately a quarter of them exhibit a neutral sentiment. A few examples of such neutral sentiment tweets are as follows: *"anyone know what's up with the geforce 500 series 580 gpx driver 275.33"* and *"hi m order is 913181 did you revise the money ? if you did .. how about the shipping ?"*. These instances appear to involve raising a complaint while simultaneously posing a question that doesn't overtly express negative sentiment. While an undercurrent of dissatisfaction is evident, the situation has not escalated to the point requiring language which explicitly expresses negative sentiment.

Next, the small number of positive tweets among the complaints were analysed. It was found they mostly involved sarcasm or where the consumer was expressing their liking for a product with the hope this could quicken the resolution process. Some of the example tweets for these scenarios are: *"hello i have a 2012 impreza and i love it . my driver seat back is broken down after 1 year and 12000 miles 32000 total 2nd owner"* and *"i love waiting at mcdonalds for 15 minutes just for some semi-good ice cream <url>"*. For tweets that were not complaints, the majority of them expressed neutral or positive sentiments.

4.1.4 Key statistics

Statistic	All Tweets	Complaints	Not Complaints	Random
Number of tweets	3449	1232	742	1475
Number of unique tweets	3395	1232	737	1427
Max tweet length (char.)	297	297	266	144
Min tweet length (char.)	1	7	6	1
Mean tweet length (char.)	77.8	96.7	70.2	65.8
Median tweet length (char.)	79.0	98.0	68.0	63.0
Standard Deviation tweet length	41.4	40.5	38.9	37.5
Total number of tokens	55169	24260	10839	20070
No. of unique tokens	7937	4031	2558	4386
Maximum tokens	57	57	55	39
Minimum tokens	1	2	1	1
Mean tokens	16.0	19.7	14.6	13.6
Median tokens	16.0	20.0	14.0	13.0
Standard Deviation for tokens	8.6	8.4	8.0	8.0
Mean punctuation count	3.4	3.9	2.7	3.3

Table 4.1: Statistics of tweets in the dataset.

Finally, examining some of the key statistical measures from the tweets dataset in Table 4.1, complaint tweets tend to exhibit a higher average tweet length, both in terms of characters (96.7) and tokens (19.7). In contrast, random tweets and replies have a token count lower by 30%, while non-complaint tweets possess an average of 25% fewer tokens than complaint tweets. This disparity may stem from individuals employing diverse linguistic expressions to express their dissatisfaction or disappointment, or to communicate a Face Threatening Act directed at the subject of the complaint.

4.2 Experiment set 1 results: Comparision of model performance

As described in the previous chapter, to compare the performance of the models a nested cross-validation approach was adopted. The cross-validation method finetuned the models using 4 learning rates, $l \, \epsilon \, [1e-5, 5e-6, 5e-5, 3e-5]$ with the best-performing model based on the F1 score being selected for the testing for each iteration of the outer loop.

4.2.1 Best predictive performance

The best-performing model was found to be BERTweet with a mean F1 score of 0.908 (sd: $\pm$0.01), accuracy of 0.934 (sd: $\pm$0.01) and ROC AUC of 0.931 (sd: $\pm$0.01) as shown in Table 4.2. Using F1 and AUC ROC scores for the assessment is more meaningful than using accuracy alone due to the class imbalance present in the dataset. BERTweet being pre-trained on a corpus of 850M tweets could be giving it an advantage in capturing the nuances of social media

posts including informal language, typographic errors, use of slang and expressive lengthening and more specific characteristics of tweets such as the use of shorter messages, abbreviations and hashtags. From the experiments, 0.00003 was the best learning rate identified from the inner loop iterations for BERTweet Base and this is used for the experiments set 2, the results for which are detailed in the next section. RoBERTa was the next best performing model with an F1 of 0.879 (sd: ±0.03), an accuracy of 0.914 (sd: ±0.02) and AUC of 0.905(sd: ± 0.02). It was followed by BERT Base and DistilBERT with F1 scores of 0.865 (sd: ±0.02) and 0.863 (sd: ±0.02) respectively.
Table 4.2 also includes the prediction metrics sourced from [18], enclosed within '[ ]', which serves as the established baseline performance for this task. Despite the variation in nested cross-validation configuration, as outlined in the preceding chapter, a level of preliminary comparison becomes feasible. In terms of F1 scores, RoBERTa shows marginal improvement, while BERT Base and ALBERT perform slightly worse. However, overall BERTweet provides the best predictive results for this task when compared to the baseline.

Model	Accuracy	Precision	Recall	F1	ROC AUC
AlBERT Base v2	0.879 [0.859]	0.845 [0.848]	0.811 [0.846]	0.827 [0.846]	0.864
⋆↑ BERTweet Base	**0.934**	**0.897**	**0.920**	**0.908**	**0.931**
BERT Base uncased	0.905 [0.88]	0.878 [0.871]	0.854 [0.873]	0.865 [0.87]	0.894
⋆↓ BERT Tiny	0.772	0.701	0.627	0.662	0.739
⋆ DistilBERT Base uncased	0.903	0.872	0.860	0.863	0.894
RoBERTa Base	0.914 [0.876]	0.886 [0.866]	0.873 [0.869]	0.879 [0.866]	0.905

Table 4.2: Mean prediction performance metrics for all models after nested cross-validation for finetuning and testing. The highest scores are in bold. ↑ is the best performing and ↓ is the worst performing model. ⋆ models are included for deep-dive analysis. Where available, numbers in '[ ]' are the results from [18].

4.2.2 Performance of smaller models

Examining the smaller models characterized as lightweight based on their architecture and parameter count, DistilBERT emerges as the top performer, achieving an F1 score of 0.863 (sd: ±0.02) and AUC of 0.894 (sd: ±0.02). ALBERT follows suit with an F1 score of 0.827 (sd: ±0.02) and AUC of (sd: ±0.04). BERT Tiny, the smallest model employed in the experiments, exhibits a notable performance gap, achieving only an F1 score of 0.662 (sd: ±0.04), which is lower by 27.6% in comparison to BERTweet. This aligns with its low accuracy and AUC scores of 0.772 (sd: ±0.02) and 0.739 (sd: ±0.03) respectively. This points to a likely and significant performance penalty from the reduced model size. However, it is expected to perform better

Model	F1	Model size	Train time	Inference time
BERTweet	*0.908*	*110M*	*87.19 s*	*0.987 s*
RoBERTa Base	-3.2%	+13.6%	-4.5%	-1.6%
BERT Base	-4.8%	+0.0%	-10.7%	+1.2%
DistilBERT	-5.0%	-40.0%	-50.7%	-43.3%
ALBERT	-9.0%	-90.0%	-40.3%	+22.8%
BERT Tiny	-27.1%	-96.0%	-86.4%	-70.2%

Table 4.3: Comparison of key metrics of the models relative to BERTweet.

in the context of a knowledge distillation teacher [41], something that has not been tested here.

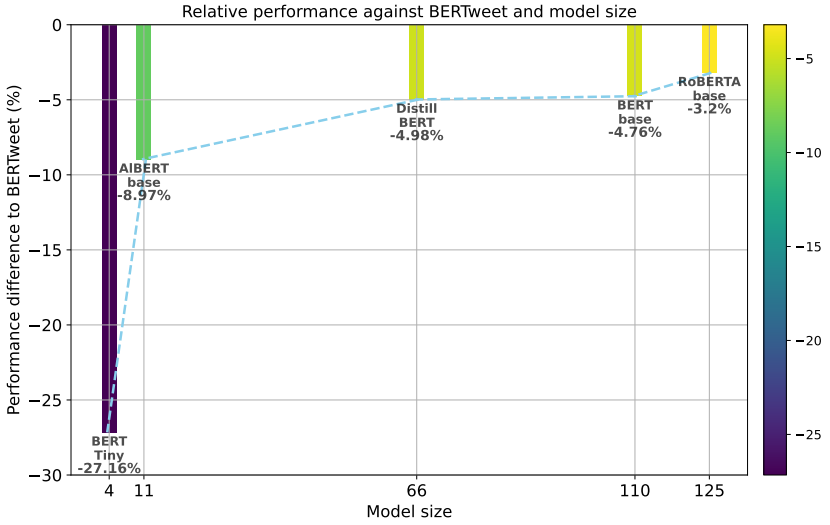


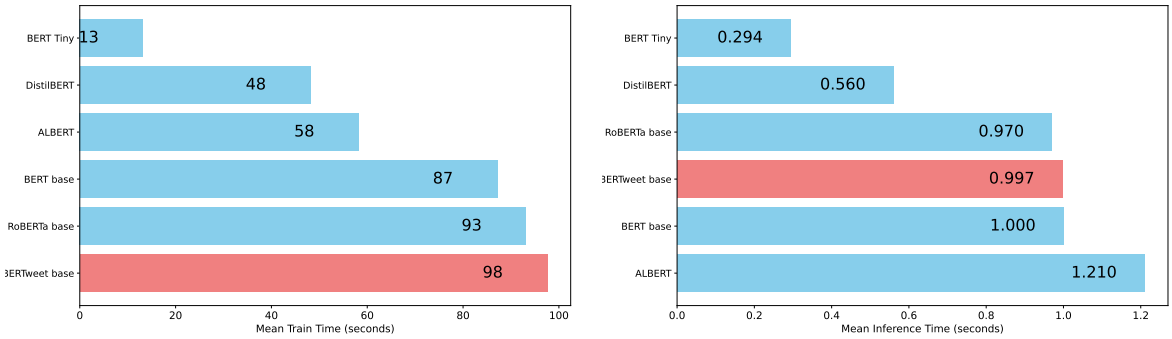
Figure 4.5: Relative performance of models against BERTweet and model sizes based on F1. BERTweet's model size is 110M.

Figure 4.5 displays the relative variance in model performance when compared to BERTweet based on F1, alongside the corresponding model sizes or the number of parameters. Table 4.3 expands on this by including the relative difference of the models compared to BERTweet in terms of model size, training time and inference time. Notably, ALBERT exhibits a relatively lower performance discrepancy of 8.9%, considering the model size is significantly smaller by 90%. DistilBERT demonstrates a performance deficit of merely 4.9%, not far off from BERT Base while having a model size lower by 44%. This could likely be due to the knowledge distillation and compression techniques applied during the pretraining phase, which could be playing a key role in enhancing DistilBERT's predictive capabilities [34].

Analyis of finetuning and inference time

Next, the time required for inference and training (finetuning) is analysed and shown in Figure 4.6. On average, the number of rows for the train, dev and test splits was 2155, 719

and 575. The training and inference were executed on a single NVIDIA RTX A4000 GPU with 16GB VRAM. BERT Tiny as expected has the lowest mean training and inference time. It is lower by over 85% and 70% for training and inference respectively as shown in Table 4.3 when compared to BERTweet. DistilBERT follows next with lower train and inference times by 51% and 43%. For ALBERT the train time is lower than BERTweet by 40%, while the inference time is slightly higher than BERT Base. ALBERT was designed to reduce the training time and memory footprint while lowering the inference time was not an explicit goal [20]. Hence these results seem to be in line with its architecture.



(a) Mean training time.

(b) Mean inference time

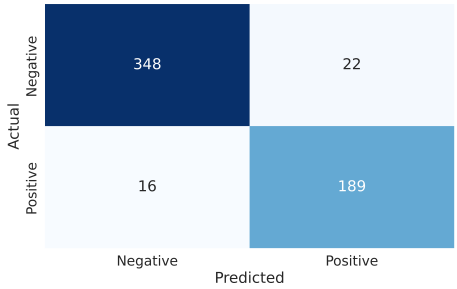
Figure 4.6: Mean time taken (in seconds) for finetuning and inference during experiments set 1. BERTweet with the best predictive model is highlighted in red.

4.2.3 Deep-dive into the results

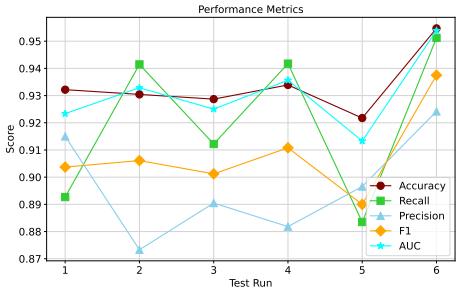
To comprehend where the models are misclassifying, deep-dive analysis is carried out on the performance of select models. The overall best-performing model, BERTweet Base, the best-performing lightweight model, DistilBERT and the worst-performing model, BERT Tiny are chosen for this exercise. The scores from Table 4.2 will be looked into in more detail along with the confusion matrix and sample misclassified tweets. The confusion matrix is based on the mean values from the 6 runs of inference carried out.

Test Run	BERTweet Base	DistilBERT	BERT Tiny
1	0.21701	0.22481	0.49180
2	0.20056	0.24783	0.50118
3	0.23015	0.26190	0.48807
4	0.23572	0.25394	0.48178
5	0.25255	0.30100	0.53076
6	0.17506	0.26712	0.48810

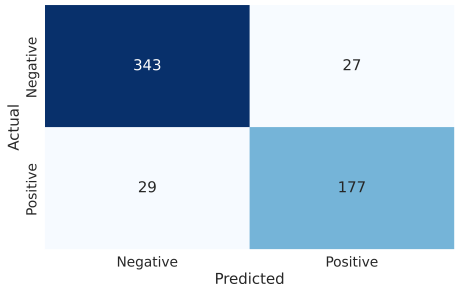
Table 4.4: Test loss from the inference phase for the 6 outer loop iterations for the 3 selected models.



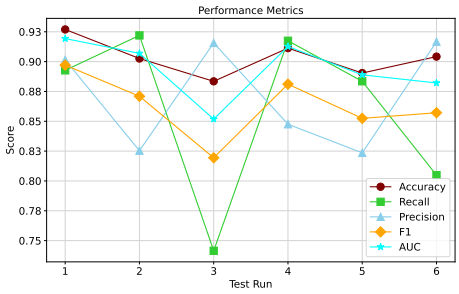
(a) BERTweet - Confusion Matrix



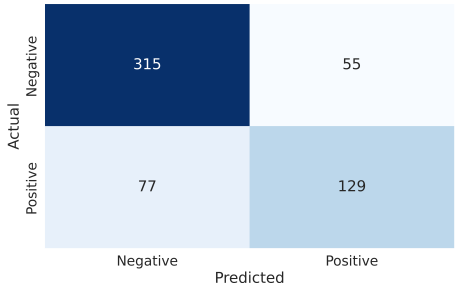
(b) BERTweet - Performance Metrics



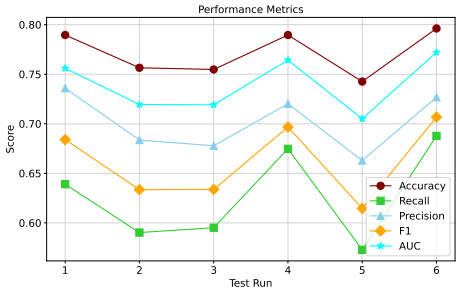
(c) DistilBERT - Confusion Matrix



(d) DistilBERT - Performance Metrics



(e) BERT Tiny - Confusion Matrix



(f) BERT Tiny - Performance Metrics

Figure 4.7: Confusion matrix and performance metrics for the 3 selected models from the inference phase. The confusion matrix is based on the mean of values from the 6 outer loop iterations and the grid can be read from left to right as true negative, false positive, false negative and true positive.

The test losses for the 6 runs are given in Table 4.4 while the graphs for the performance scores from inference are illustrated in Figure 4.7. As expected, the test loss is lower for most the of 6 runs for BERTweet Base compared to the other two models. DistilBERT Base shows losses that are closer to BERT Base in 2 runs but otherwise are generally higher. There is a considerable difference for BERT Tiny which reflects in the predictive performance scores discussed earlier.

Referring to the metrics in Figure 4.7, BERT Tiny exhibits a pattern that differs from the other 2 models. Based on the correct and incorrect classifications for all the inference

executions (detailed in Appendix A.3.1), it appears that within each test run, the change in the misclassified complaint and non-complaint tweets move in the same direction. Another point to note is that BERTweet and DistilBERT seem to have relatively high fluctuations in recall and precision, likely due to inherent variations in the data splits from the cross-folds for each run. Considering the class imbalance, the analysis includes evaluating the ROC AUC scores. BERTweet demonstrates superior discrimination capability compared to the other two models, as reflected in its AUC score while exhibiting the least variability with a standard deviation of 0.01. DistilBERT has the next best score but displays slightly greater variance with a standard deviation of 0.02. In contrast, BERT Tiny showcases the lowest AUC scores and the highest standard deviation at 0.03.

Figure 4.7 also shows the confusion matrices for the 3 models based on the mean of the classifications from the 6 test runs. The ratio of accurately categorized non-complaint tweets (true negatives) to complaint tweets (true positives) is comparable between BERTweet and DistilBERT, but it's lower for BERT Tiny. All models appear to face greater challenges in accurately classifying complaint tweets compared to non-complaint tweets, although BERTweet has the best recall scores overall. Even though there is an inherent class imbalance in the dataset, the confusion matrix and the metrics presented still offer valuable insights into the predictive capabilities of the models

Next, an error analysis is conducted on the tweets that were misclassified. Sample tweets that were inaccurately classified by the three models are presented in Table 4.5. Examining tweets that were wrongly labelled as non-complaints several observations are discussed. BERTweet seems to have challenges in identifying acts of complaining which include the usage of very concise text (example 1) or very limited context (example 2). DistilBERT is unable to identify the use of sarcasm (example 5) and the usage of rhetorical questions (example 6). Notably, BERT Tiny struggles even when tweets exhibit multiple indicators of complaining, as seen in examples 9 and 10. Regarding tweets that were erroneously classified as complaints, the potential reasons are more difficult to speculate on. There are tweets containing words commonly associated with grievances, like 'waiting' and 'issue' (examples 3 and 4 for BERTweet) which is possibly confusing the model. DistilBERT misclassifies tweets that narrate negative experiences rather than expressing complaints (example 7) as well as instances where the text represents an intermediate segment of a conversation (example 8). In the case of the latter, while there is a hint of complaining it is not an act of complaining in this context. BERT Tiny appears to get confused when encountering text containing questions that carry no hint of complaining and are about mundane topics (examples 11 and 12). There possibly could be improvements to be found if more examples of such scenarios are included in the data used for finetuning, especially for the BERTweet and DistilBERT models.

No.	Tweet
	BERTweet
1	worst *
2	. <user> <user> cuda driver lion update , please ? <url>
3	i had entered in giveway contest for oneplus 5t lava . so waiting for the result
4	i just checked again and i'm not having the same issues i had earlier . thanks for the help
	DistilBERT
5	just want to thank <user> <user> for ignoring me for three days xxxx
6	is this really what you call a large milkshake <user> <url>
7	shout out to the social media team <user> <user> whilst i get the frustration , there's never a time people should be insulting or rude when tweeting . these good people responding are employee's just tyring to help . #heathrowairport #heathrow #britishairways
8	no , see screenshot . it's the app . i got it straight off the playstore on android .
	BERT Tiny
9	whyyyy is your wifi so slow <user> ?
10	my 2nd visit to kfc chadwell heath after the last fiasco . this time they had no gravy or corn amp ; forgot my chips #fail
11	is it possible to integrate my medium account on my personal website with your api ?
12	thanks for your response . what is the twitter handle for your care centre in delhi , india ?

Table 4.5: Sample tweets which have been misclassified by the 3 selected models. Tweets in the lighter shade of grey are misclassified as complaints while the rest are misclassified as not complaints.

Train \ Test	Food	Appr.	Cars	Retail	Srvcs.	Softw.	Trans.	Elect.	Other
Food	-	0.515 0.714	0.522 0.847	0.528 0.774	0.532 0.765	**0.534** 0.797	0.517 0.718	0.529 0.762	0.510 **0.855**
Appr.	**0.854 0.918**	-	0.775 0.901	0.828 0.864	0.807 0.860	0.852 0.880	0.801 0.840	0.795 0.856	0.830 0.916
Cars	0.500 0.840	0.500 0.713	-	0.500 0.765	0.500 0.757	0.500 0.789	0.500 0.717	0.500 0.755	0.500 **0.852**
Retail	**0.798 0.889**	0.706 0.788	0.715 0.888	-	0.698 0.812	0.752 0.869	0.698 0.787	0.701 0.821	0.666 0.889
Srvcs.	0.820 0.804	0.829 0.821	0.781 0.860	0.812 0.817	-	0.834 0.856	0.802 0.811	0.824 0.852	**0.885 0.913**
Softw.	**0.789** 0.848	0.761 0.809	0.747 0.901	0.786 0.852	0.738 0.838	-	0.736 0.806	0.757 0.845	0.748 **0.906**
Trans.	0.853 0.859	0.813 0.819	0.800 0.898	0.833 0.856	0.807 0.854	0.840 0.876	-	0.777 0.825	**0.871 0.915**
Elect.	0.825 0.842	0.834 0.839	0.813 0.896	0.847 0.856	0.825 0.876	**0.853** 0.891	0.790 0.812	-	0.843 **0.919**
Other	0.500 0.840	0.500 0.713	0.500 **0.845**	0.500 0.765	0.500 0.757	0.500 0.789	0.500 0.717	0.500 0.755	-
All	0.870 0.869	0.856 0.871	0.837 0.903	0.879 0.889	0.851 0.874	0.882 0.905	0.824 0.829	0.838 0.848	**0.908 0.951**

Table 4.6: ROC-AUC and F1 scores for the cross-domain experiments are recorded here. The rows show the domain used for finetuning while the columns represent the domains used for testing. The last row shows the scores where the full data except the corresponding test domain was used for finetuning. Best scores where applicable are highlighted in bold.

4.3 Experiment set 2 results: Cross-domain results

What follows are the results from the cross-domain experiments. To recap, these experiments aim to evaluate the behaviour of the top-performing model from experiments set 1 in the context of smaller datasets and imbalanced class distributions. For these experiments, the chosen model is BERTweet Base, utilizing the optimal learning rate hyperparameter of 0.00003. The outcomes are detailed in Table 4.6, focusing on the assessment performed using ROC-AUC and F1 scores. The AUC metric is included due to its appropriateness for datasets with class imbalance. The metrics are reported as the mean of 3 test runs based on a stratified cross-fold split of the data. For a comprehensive breakdown of classes within each domain, please refer to Appendix A.1.

The best performance is where all the data except the 'Other' domain tweets is used for finetuning and the 'Other' domain for testing. It has a relatively high discriminative ability represented by an AUC score of 0.908. Moreover, the achieved F1 score of 0.951 indicates a good balance between its precision and recall as well. The use of diverse domains for finetuning likely has a positive impact on how the models learn for the downstream task. In general, the predictive performance of the experiments where all data except one domain is used for finetuning is consistently higher when compared to using just a single domain for finetuning.

For single-domain finetuning cases, some instances exhibit performance levels approaching those attained through finetuning with the entire dataset, even though the volume of training data is significantly lower. Domains such as 'Apparel', 'Services', 'Transport' and 'Electronics' have AUC and F1 scores of over 0.8 despite the volume of data being on average only 10% of the 'All' data volume. However, it's also important to highlight that finetuning with single domains with extremely limited data for finetuning results in AUC scores around or close to 0.5. This indicates that the model's predictive ability is akin to random chance in these cases. These include the domains of Cars', 'Food & beverage', and 'Others' with only 4% of the 'All' data volume on average. This is similar to the observations from [18]. In the overall context, the notable performance of 'Food' and 'Other' as testing domains should be approached with caution, given the limited amount of inference data available.

The best score achieved by the previous baseline [18] using BERT Base was 0.882 when employing all domains for finetuning and the 'Other' domain for testing. While BERTweet's best performance is for the same combination, the F1 score is significantly higher at 0.951. This notable difference in performance can likely be attributed to the advantages previously outlined for BERTweet Base, which is pre-trained on Twitter data. The predictive performance when finetuning with domains with very low tweet volumes is also higher for BERTweet. For the 'Cars' domain, which uses the lowest volume of data, BERT Base scores an average F1 of 0.623 while BERTweet achieves a substantially higher F1 of 0.774.

Continuing to explore the relationship between the volume of data and the performance,

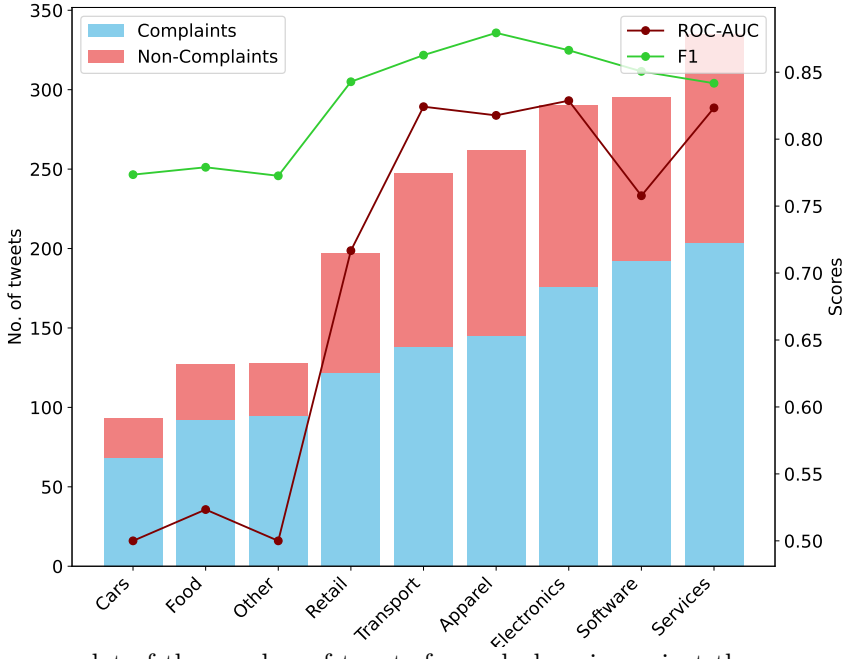


Figure 4.8: Shows a plot of the number of tweets for each domain against the average performance when that domain is used for finetuning.

Figure 4.8 illustrates the impact of the number of instances or tweets utilized during the finetuning process on the average testing performance. Additionally, it shows the breakdown of complaint and non-complaint tweets. At a macro level, it appears the predictive performance improves as the volume of data available for finetuning increases. However, specific domains like 'Other', 'Apparel', and 'Software' exhibit deviations from this trend. Inherent characteristics of tweets within these domains could be influencing their performance deviations. All domains predominantly feature complaints as the dominant class. However, no distinct pattern emerges from how the variations in the class imbalance impact performance in this situation.

Chapter 5

Conclusions

With the advancement of technology and the widespread adoption of social media, the anticipated response times for businesses to address complaints have significantly reduced. Furthermore, this evolution has introduced numerous online avenues through which customers can seek assistance or voice their grievances. With the substantial visibility offered by these platforms, organisations are exposed to the effects of negative online word of mouth. Consequently, the implementation of tools to automatically detect complaints has many benefits for organisations.

Building upon earlier research in this field, this study has led to a few key conclusions. First and foremost, it's apparent that BERTweet outperforms the other assessed transformer models in identifying complaints posted on Twitter. Notably, BERTweet shows comparatively good performance even when dealing with limited finetuning data. However, caution is advised when generalizing this finding to other platforms like Facebook, given the constraints of the pre-training data (limited to Twitter data). Additionally, a significant takeaway is that organizations constrained by memory and computational resources can effectively opt for smaller models like DistilBERT without significantly compromising predictive performance.

In conclusion, there exist several promising paths for further research within this domain. Analyzing the performance of similar models on alternative social media platforms like Facebook could yield deeper insights into the models' capabilities on linguistically diverse datasets. Another avenue involves evaluating the multi-modal approach with the BERTweet model and whether the inclusion of additional linguistic cues could enhance its performance. Lastly, considering the ongoing advancements in generative models, exploring prompting with solutions that use state-of-the-art models like GPT 3.5 / 4 could have great value.

Bibliography

[1] Artstein, R., and Poesio, M. Inter-Coder Agreement for Computational Linguistics. *Computational linguistics - Association for Computational Linguistics 34*, 4 (2008), 555–596.

[2] Balaji, M. S., Jha, S., and Royne, M. B. Customer e-complaining behaviours using social media. *The Service industries journal 35*, 11-12 (2015), 633–654.

[3] Becherer, R. C., and Richard, L. M. Self-Monitoring as a Moderating Variable in Consumer Behavior. *The Journal of consumer research 5*, 3 (1978), 159–162.

[4] Bhargava, P., Drozd, A., and Rogers, A. Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics, Oct. 2021.

[5] Boxer, D. Social distance and speech behavior: The case of indirect complaints. *Journal of pragmatics 19*, 2 (1993), 103–125.

[6] Brown, P., and Levinson, Stephen C. *Politeness : some universals in language usage*. Studies in interactional sociolinguistics ; 4. Cambridge University Press, Cambridge [Cambridgeshire] ; New York, 1987.

[7] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners, July 2020.

[8] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, Dec. 2014.

[9] Coussement, K., and Van den Poel, D. Improving customer complaint management by automatic email classification using linguistic style features as predictors. *DECISION SUPPORT SYSTEMS 44*, 4 (2008), 870–882.

[10] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019.

[11] Faucher, K. X. *Social Capital Online: alienation and accumulation.* CDSMS (Series). University of Westminster Press, London, London, England, 2018.

[12] Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer, New York, NY, 2009.

[13] Hennig-Thurau, T., Gwinner, K. P., Walsh, G., and Gremler, D. D. Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of interactive marketing 18*, 1 (2004), 38–52.

[14] Hirschberg, J., and Manning, C. D. Advances in natural language processing. *Science 349*, 6245 (July 2015), 261–266.

[15] Hochreiter, S., and Schmidhuber, J. Long Short-Term Memory. *Neural Computation 9*, 8 (Nov. 1997), 1735–1780.

[16] Howard, J., and Ruder, S. Universal Language Model Fine-tuning for Text Classification, May 2018.

[17] Istanbulluoglu, D. Complaint handling on social media: The impact of multiple response times on consumer satisfaction. *Computers in Human Behavior 74* (Sept. 2017), 72–82.

[18] Jin, M., and Aletras, N. Complaint Identification in Social Media with Transformer Networks, Oct. 2020.

[19] Jin, M., and Aletras, N. Modeling the Severity of Complaints in Social Media, Mar. 2021.

[20] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, Feb. 2020.

[21] Lau, G. T., and Ng, S. Individual and Situational Factors Influencing Negative Word-of-Mouth Behaviour. *Canadian Journal of Administrative Sciences / Revue Canadienne des Sciences de l'Administration 18*, 3 (2001), 163–178.

[22] Liang, C.-Y., Guo, L., Xia, Z.-J., Nie, F.-G., Li, X.-X., Su, L., and Yang, Z.-Y. Dictionary-based text categorization of chemical web pages. *Information Processing & Management 42*, 4 (2006), 1017–1029.

[23] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019.

[24] Lui, M., and Baldwin, T. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the ACL 2012 System Demonstrations* (Jeju Island, Korea, July 2012), Association for Computational Linguistics, pp. 25–30.

[25] Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer Sentinel Mixture Models, Sept. 2016.

[26] Nguyen, D. Q., Vu, T., and Nguyen, A. T. BERTweet: A pre-trained language model for English Tweets, Oct. 2020.

[27] Olshtain, E., and Weinbach, L. Complaints: A study of speech act behavior among native and non-native speakers of Hebrew. 195–208.

[28] OpenAI. GPT-4 Technical Report, Mar. 2023.

[29] Pfeffer, J., Zorbach, T., and Carley, K. M. Understanding online firestorms: Negative word-of-mouth dynamics in social media networks. *Journal of Marketing Communications 20*, 1-2 (Mar. 2014), 117–128.

[30] Preotiuc-Pietro, D., Gaman, M., and Aletras, N. Automatically Identifying Complaints in Social Media, June 2019.

[31] Pérez, J. M., Giudici, J. C., and Luque, F. pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks, June 2021.

[32] Qiu, L., Lin, H., Ramsay, J., and Yang, F. You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality 46*, 6 (Dec. 2012), 710–718.

[33] Rook, D. W., and Fisher, R. J. Normative Influences on Impulsive Buying Behavior. *The Journal of consumer research 22*, 3 (1995), 305–313.

[34] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, Feb. 2020.

[35] Shane-Simpson, C., Manago, A., Gaggi, N., and Gillespie-Lynch, K. Why do college students prefer Facebook, Twitter, or Instagram? Site affordances, tensions between privacy and self-expression, and implications for social capital. *Computers in Human Behavior 86* (Sept. 2018), 276–288.

[36] Sharma, P., Marshall, R., Alan Reday, P., and Na, W. Complainers versus non-complainers: a multi-national investigation of individual and situational influences on customer complaint behaviour. *Journal of marketing management 26*, 1-2 (2010), 163–180.

[37] Sparks, B. A., and Browning, V. Complaining in Cyberspace: The Motives and Forms of Hotel Guests' Complaints Online. *Journal of hospitality marketing & management 19*, 7 (2010), 797–818.

[38] Sun, S., Gao, Y., and Rui, H. Does Active Service Intervention Drive More Complaints on Social Media? The Roles of Service Quality and Awareness. *Journal of Management Information Systems 38*, 3 (July 2021), 579–611.

[39] Tripp, T. M., and Gregoire, Y. When unhappy customers strike back on the Internet. *MIT Sloan management review 52*, 3 (2011), 37.

[40] Trosborg, A. *Interlanguage pragmatics: Requests, complaints, and apologies*, vol. 7. Walter de Gruyter, 2011.

[41] Turc, I., Chang, M.-W., Lee, K., and Toutanova, K. Well-Read Students Learn Better: On the Importance of Pre-training Compact Models, Sept. 2019.

[42] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention Is All You Need, Aug. 2023.

[43] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. HuggingFace's Transformers: State-of-the-art Natural Language Processing, July 2020.

Appendices

Appendix A

Other supporting analysis and graphs

A.1 Breakdown of tweets in full dataset

Domains	Complaints	Non-Complaints	Total Tweets
Apparel	145 (55.3%)	117 (44.7%)	262 (7.6%)
Cars	68 (73.1%)	25 (26.9%)	93 (2.7%)
Electronics	176 (60.7%)	114 (39.3%)	290 (8.4%)
Food & Beverage	92 (72.4%)	35 (27.6%)	127 (3.7%)
Other	95 (74.2%)	33 (25.8%)	128 (3.7%)
Retail	122 (61.9%)	75 (38.1%)	197 (5.7%)
Services	204 (60.9%)	131 (39.1%)	335 (9.7%)
Software & Online Services	192 (65.1%)	103 (34.9%)	295 (8.6%)
Transport	138 (55.9%)	109 (44.1%)	247 (7.2%)
Sub-total	**1232 (62.4%)**	**742 (37.6%)**	**1974**
Random Reply	0 (0%)	739 (100%)	739 (21.4%)
Random Tweet	0 (0%)	736 (100%)	736 (21.3%)
Total	**1232 (35.7%)**	**2217 (64.3%)**	**3449**

Table A.1: The nine domains and the distribution of tweets that are complaints and those that are not from the latest version of the dataset available in the public domain and for the experiments. Additionally, the table includes the number of random tweets and replies introduced into the dataset by the authors for a more proportionate representation of the classes.

A.2 Sample data from dataset

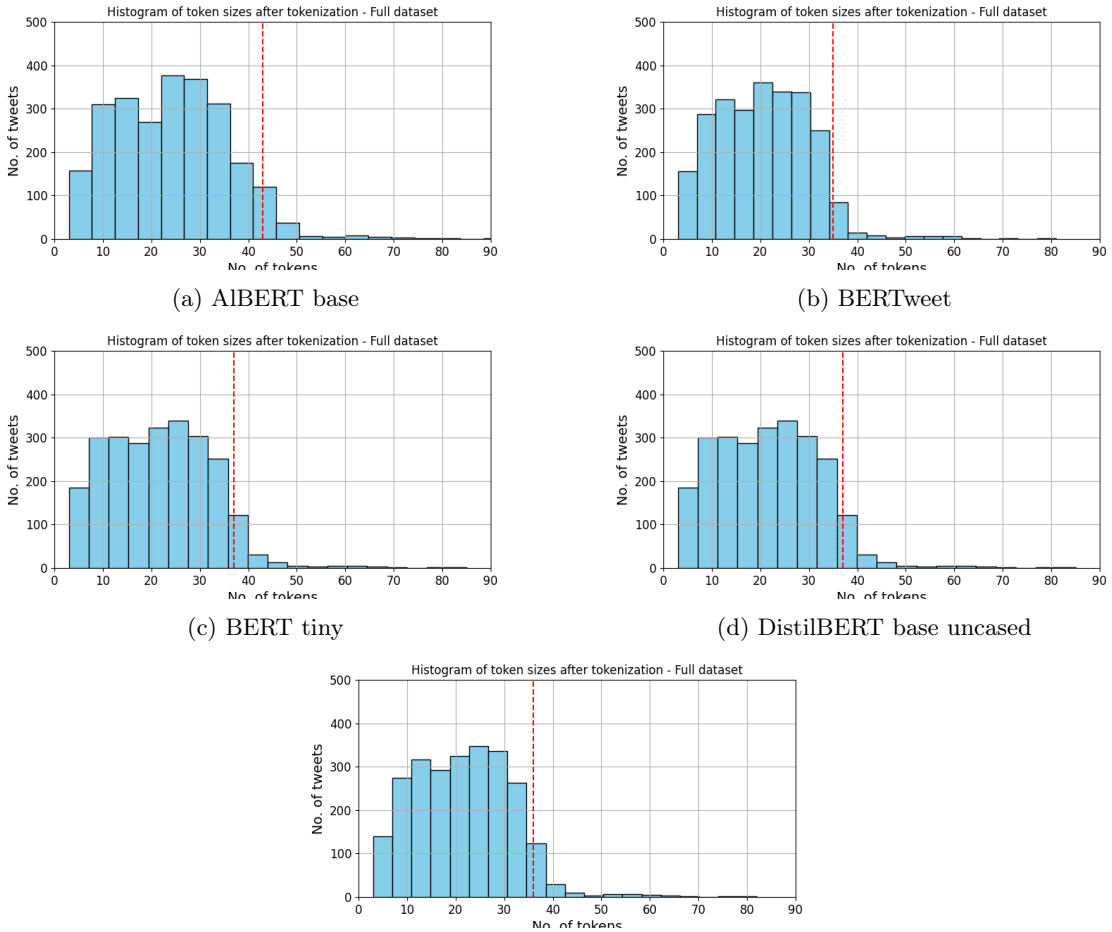
id	text	binarylabel	multilabel	domain
1.20E+17	asus g60 series . bought it to play games but guess not bf3	1	1	electronics
4.88E+17	love this trimmer ! making the sidewalk look sharp <url>	0	0	other

Table A.2: Sample data from [19].

The `binarylabel` represents the label for complaints with 1 indicating the tweet is a complaint. Columns `id` and `multilabel` are not used for the experiments.

A.3 Token distribution after tokenization

The graphs in Figure A.1 show the distribution of token size for tweets from the full dataset for each of the models based on the tokenizer they use. The graph for BERT base uncased is in Chapter 3, Figure 3.2.



(a) AlBERT base

(b) BERTweet

(c) BERT tiny

(d) DistilBERT base uncased

(e) RoBERTa base

Figure A.1: The token count distribution for the full dataset of 3,449 tweets for all models.

A.3.1 Confusion matrix from every test run

The elements of the matrix in left-to-right order are true negative, false positive, false negative and true positive.

Test run	BERTweet Base	DistilBERT Base	BERT Tiny
1	[353 17] [22 183]	[350 20] [22 183]	[323 47] [74 131]
2	[342 28] [12 193]	[330 40] [16 189]	[314 56] [84 121]
3	[347 23] [18 187]	[356 14] [53 152]	[312 58] [83 122]
4	[343 26] [12 194]	[335 34] [17 189]	[315 54] [67 139]
5	[348 21] [24 182]	[330 39] [24 182]	[309 60] [88 118]
6	[353 16] [10 195]	[354 15] [40 165]	[316 53] [64 141]

Table A.3: Confusion matrix from every test run for the experiments set 1.

Appendix B

Other references

B.1 Code Repository

The code (Jupyter notebook) used for the experiments in this report and the results from the finetuning and testing are available at `https://github.com/nsmathew/transformers_complaints`.

B.2 References for models used in experiment sets 1 and 2

Model	Model Documentation
AlBERT base	`https://huggingface.co/albert-base-v2`
BERT base (uncased)	`https://huggingface.co/bert-base-uncased`
BERT Tiny	`https://huggingface.co/prajjwal1/bert-tiny`
BERTweet base	`https://huggingface.co/vinai/bertweet-base`
DistilBERT base (uncased)	`https://huggingface.co/distilbert-base-uncased`
MobileBERT (uncased)	`https://huggingface.co/google/mobilebert-uncased`
RoBERTa base	`https://huggingface.co/roberta-base`

Table B.1: The transformer models used for the experiments and links to their documentation.

B.3 References for other models used

Model	Model Documentation
BERTweet base sentiment analysis (pysentimiento)	`https://huggingface.co/finiteautomata/bertweet-base-sentiment-analysis`

Table B.2: The other models used in the chapter on results and their documentation.

B.4 Evaluation metrics references

Metric	Metric Documentation
Accuracy	https://huggingface.co/spaces/evaluate-metric/accuracy
F1	https://huggingface.co/spaces/evaluate-metric/f1
Precision	https://huggingface.co/spaces/evaluate-metric/precision
Recall	https://huggingface.co/spaces/evaluate-metric/recall
ROC AUC	https://huggingface.co/spaces/evaluate-metric/roc_auc
Matthews correlation coefficient	https://huggingface.co/spaces/evaluate-metric/matthews_correlation
Confusion Matrix	https://huggingface.co/spaces/BucketHeadP65/confusion_matrix

Table B.3: The metrics used for evaluating the performance of the experiments and links to their documentation.

B.5 Additional information on environment used

- **CPU Model:** Intel® Xeon® Gold 5315Y Processor
- **Operating System:** Ubuntu 20.04.5 LTS