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 across our lives and the active enablement of such platforms by organisations for user engagement, it has become a common medium for users to raise complaints. With such complaints being publicly visible, it is imperative 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.

In this paper, the use of these approaches is extended by assessing variants of BERT, including BERTweet, which is pre-trained on twitter data and 'lightweight' models such as DistillBERT, MobileBERT, BERT tiny which are meant to reduce the time required for fine-tuning as well as inference. The performance of these 'lightweight' models is compared with the traditional transformer models including BERT, RoBERTa, BERTweet for this particular text classification task. The dataset used consists of anonymised and annotated(complaint or not) Twitter data utilized in previous research and currently available in the public domain. In addition, the act of complaining and the nature of complaints are analysed from a linguistic perspective along with discussions on state-of-the-art approaches for such NLP tasks.

Update with high level results

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 [17]. 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 of the complaint. 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 and specifically social media, the impact of negative 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 [25]. Customers are able to express their complaints directly, conveniently, and with enhanced effectiveness to organisations through multiple social media channels and platforms [2].

By examining instances of complaints on social media and specifically Twitter, we find them 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, we find the objective being 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 customer service* and hashtags like *#pissed*, *#useless* and *#worstbrand*. One could also argue that the

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.

second tweet encompasses elements of a Face Threatening Act, given that implicating 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 a more detailed 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 [18].

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 [14]. Latent Semantic Indexing based on Singular Value Decomposition along with linguistic style features has been utilised to classify emails as complaints or not [8]. In recent years, we have seen the use of various Machine learning 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 [18] on Twitter feeds. The use of more advanced approaches using transformer networks has shown to have better results as explored by [12]. As part of this paper, the use of the BERT and its many variants, including that of recently created lightweight versions will be assessed further on a publicly available Twitter dataset.

1.2 Aims and Objectives

TO UPDATE

Chapter 2

Literature Survey

2.1 The act of complaining

As per [17], 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 [23].

The authors of [17] further delve into the intentions of the speaker in making the complaint. They argue this is carried out with either the hope of repair of 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 additionally 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, 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, tweets which fall into the category of indirect complaints are also included in the dataset.

Analysing deeper into which types of customers complain more, [22] have looked at how personality traits like impulsivity and self-monitoring impact customer complaining behaviour. *Impulsivity*, as defined by [20], 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. [3] defines *self-monitoring* 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, [22] 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. However, these effects are more pronounced in situations where the level of dissatisfaction is high.

2.2 Complaining online

The act of complaining exists online in various forms and with varying degrees of intensity and this prevalence lead to the emergence of third-party organisations that provide online channels for customers' ease and convenience [25]. 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 [25].

Of the 431 online complaints assessed by [25], 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 [10]. This type of communication can take place on various platforms, ranging from official company-sponsored sites to unaffiliated blogs. The Internet offers consumers a convenient and anonymous platform to express negative word-of-mouth by sharing their viewpoints and complaints with others. Among the different forms of EWOM, consumer reviews are particularly noteworthy, as they provide valuable insights about products, whether positive or negative [23]. Such Negative electronic WOM (EWOM), can significantly damage a brand's reputation and influence potential customers to seek alternative products or services.

Technology provides an accessible channel that allows consumers to complain with significant ease, making it available to anyone with internet access, even those who may be hesitant to complain directly to the company [23]. The reviews and comments posted by consumers online can hold considerable influence over decisions made by other fellow consumers. From an organisation's perspective, the use of online complaining by consumers has some indirect negative consequences as well. The potential experience and knowledge frontline personnel could gain from addressing the complaints directly are lost and this has long-term implications for the organisation. The study by [23] on online complaining in the hospitality industry, corroborates the double deviation theory touched upon earlier. Most complaints reviewed involved the individual complaining online after having failed to receive a satisfactory resolution from the hotel staff. Another key finding was the altruistic nature of the complaints, intending to warn other potential customers of the problems. The nature of complaints points to a sense of unfairness being experienced by the guests due to their initial complaints being inadequately addressed and in some cases, combined with a lack of empathy from the staff.

2.3 Complaining in social media

TO UPDATE

2.4 Self-expression on Twitter

TO UPDATE

2.5 Transformers

TO UPDATE

2.6 Ongoing research

TO UPDATE

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. Twitter provides a good representation of social media text due to the direct connection consumers have with organisations and brands as well as the ability to express oneself [18]. **Add content on why Twitter**

The data set created by [18] and further used by [12] 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 [13] 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 based on 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.

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

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 [18].

3.2.2 Data Extraction

The data was extracted from Twitter via the Twitter API[2]. The latest 3,200 tweets at the time of the collection exercise were extracted and the original tweets to which the customer service handles 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 [15]. 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) [18].

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

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

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 [17], 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 on a single GPU.

3.3.1 Hardware

- **CPU Count:** 8
- **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]. 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.

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 [9] is based on the modern transformers network architecture [27]. Using BERT for the text classification task has several advantages over previous dominant methods such as Gated Recurrent Units GRU) [7] or Long Short Term Memory (LSTM) [11]

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

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

networks. Although tweets tend to be made up of short texts, the ability to capture long-term dependencies is still useful for better 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 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 (we use only 3,449 tweets 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 and BERT base with 110 million parameters to lightweight variants such as DistilBERT base, MobileBERT and others 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 in relation to the model size. **Add content on the impact of layers and parameters on model performance**

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
MobileBERT (uncased)	25.3M	WordPiece	30,522
ALBERT base	11M	SentencePiece	30,000
BERT Tiny	4.4M	WordPiece	30,522

Table 3.4: The transformer models used for the experiments along with 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, BERT base, ALBERT and BERT Tiny. For Bertweet it is from [16], MobileBERT from [24] and DistilBERT from [21].

3.5 Data tokenisation

The tokenization process is required to be applied to input data for it to be prepared appropriately 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 with each having advantages and shortcomings. The methods used by each of the models in the scope of the experiments are briefly described below and shown in Table 3.4.

3.5.1 Tokenization methods

Byte-Pair Encoding or BPE: BPE works by iteratively combining the most frequently occurring pairs of characters or subwords within a corpus until a predefined vocabulary size is reached or after reaching a maximum number of iterations. The vocabulary will consist of a set of subwords, which can include characters, character sequences, or partial words.

Byte-level BPE: This approach works similarly to BPE but operates at byte level, treating each byte of a text as a token and merging the most frequent pairs of bytes in a text corpus.

WordPiece: WordPiece segments text into subword entities by identifying probable splits that optimize likelihood within the training data. The tokenizer may combine often co-occurring subword pairs to create new subword elements. This sequence of merging is carried out until a predefined vocabulary size is reached. The resulting vocabulary encompasses these subword components, which may be entire words or word fragments.

SentencePiece: SentencePiece treats the entire text as a single stream of text and splits the input text into subword units with whitespace also handled as a normal symbol. While training, SentencePiece generates a vocabulary of subword units based on the given text corpus. Subword units are selected such that the frequent patterns in the text are captured.

The tokenizer provided by the `transformers` library is used for tokenizing 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.

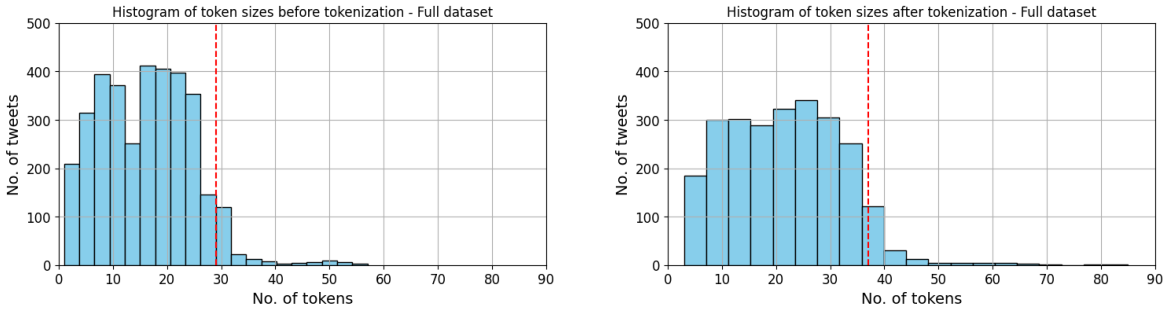
[6] `https://huggingface.co/docs/transformers/v4.21.3/en/model_doc/bert#transformers.BertTokenizer`

[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`

3.5.2 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, `padding` controls the type of padding and `truncation` allows to truncate input to a pre-determined number of tokens.



(a) Before tokenization

(b) After tokenization

Figure 3.1: 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 [18, 13] before applying the model-specific tokenization is shown in Figure 3.1a. Over 95% of the tweets have 29 tokens or less. Using the BertTokenizer as an example, from Figure 3.1b 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.

3.5.3 Tokenisation example

An example tweet from the input data is shown in **A**. The pre-processing applied by [18, 13] 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.

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

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: 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 [18]. 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 [18]

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 6 stratified splits capturing sufficient variability in the input dataset.

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.

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 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 performance metrics, precision, recall, accuracy, F1, and 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 performance

To evaluate the performance of finetuned models using a significantly small dataset with additional constraints of class imbalance, cross-domain experiments are conducted. 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

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

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 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 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 [18] and further enhanced with complaints severity type annotation by [13]. 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 [18] and further enhanced with the labels for complaint severity by [13]. 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

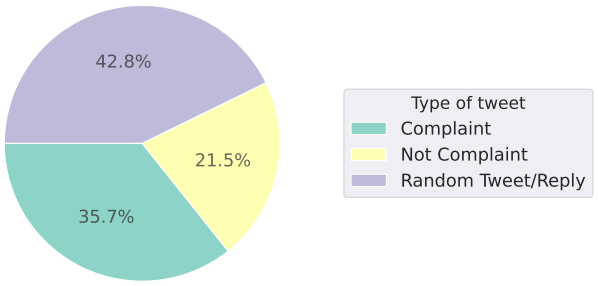
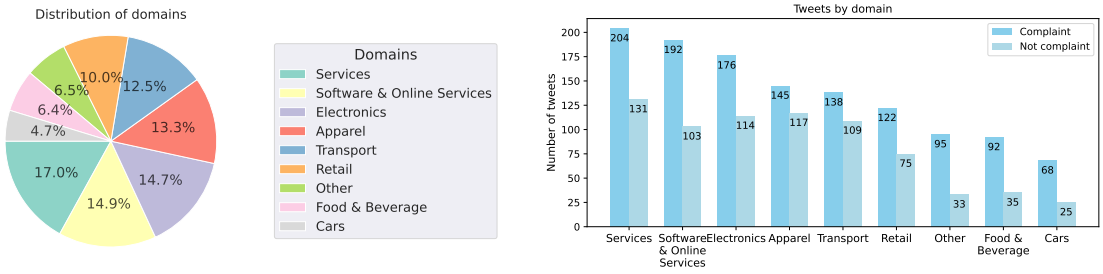


Figure 4.1: Illustrates the distribution of tweets categorised as 'complaints' and 'not complaints', with random 'tweets / replies' shown separately.

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 [18] 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.

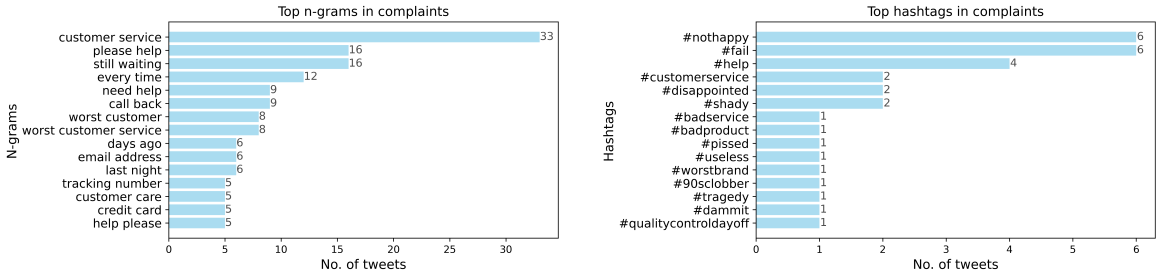


(a) Proportion of each domain

(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

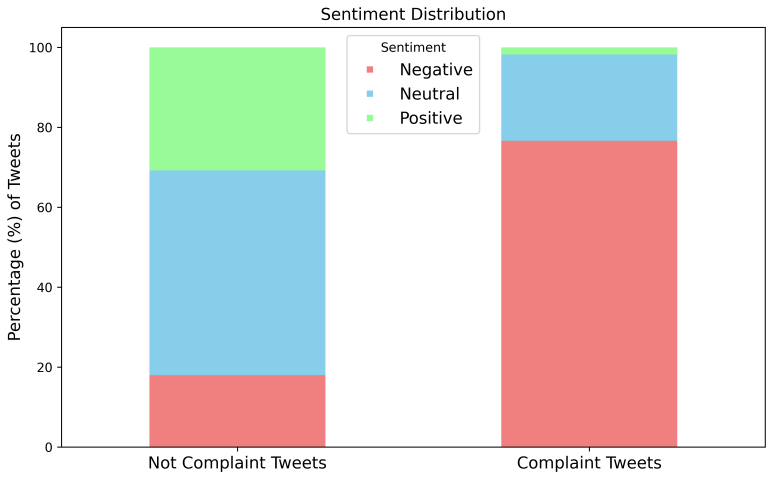


Figure 4.4: Distribution of positive, negative and neutral sentiments in the tweets.

The act of complaining typically involves conveying a sense of negativity. Sentiment analysis was conducted on the dataset using pysentimiento's library [19] 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 couple of 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. For tweets that are not complaints, the majority of them express neutral or positive sentiments. **CHECK ON POSITIVE TWEET IN COMPLAINTS**

4.1.4 Key statistics

Finally, examining some of the key statistical measures from the tweets dataset in Table 4.1, we observe that 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

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.

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) and accuracy of 0.934 (sd: $\pm$0.01) as shown in Table 4.2. BERTweet being pre-trained on a corpus of 850M tweets seems to give it a significant 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. RoBERTa was the next best performing model with an F1 of 0.879 (sd: $\pm$0.03) and an accuracy of 0.914 (sd: $\pm$0.02). It was followed by BERT base and DistilBERT with F1 scores of 0.865 (sd: $\pm$0.02) and 0.863 (sd: $\pm$0.02) respectively.

Table 4.2 also includes the prediction metrics sourced from [12], 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
MobileBERT uncased	0.887	0.843	0.843	0.841	0.877
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. Where available, numbers in '[ ]' are the results from [12].

4.2.2 Performance of 'lightweight' models

Examining the 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). MobileBERT and ALBERT follow suit with F1 scores of 0.841 (sd: ±0.02) and 0.827 (sd: ±0.05) respectively. 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 points to a likely and significant performance penalty from the reduced model. However, it is expected to perform better in the context of a knowledge distillation teacher [26], something that has not been tested here.

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 and MobileBERT exhibit relatively lower performance discrepancies of 8.9% and 7.4%, considering the model sizes are significantly smaller by 90% and 77% respectively. 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

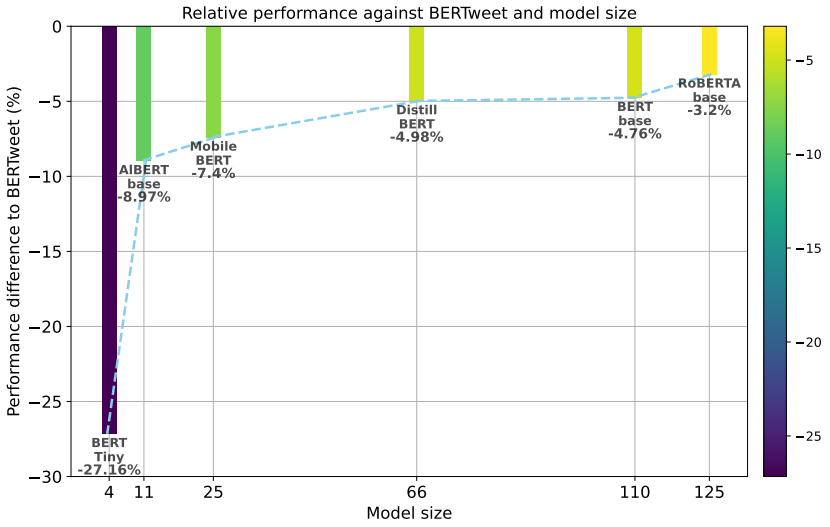


Figure 4.5: Relative performance of models against BERTweet and model sizes based on F1. BERTweet's model size is 110M.

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%
MobileBERT	-7.4%	-77.0%	+122.7%	+86.4%
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.

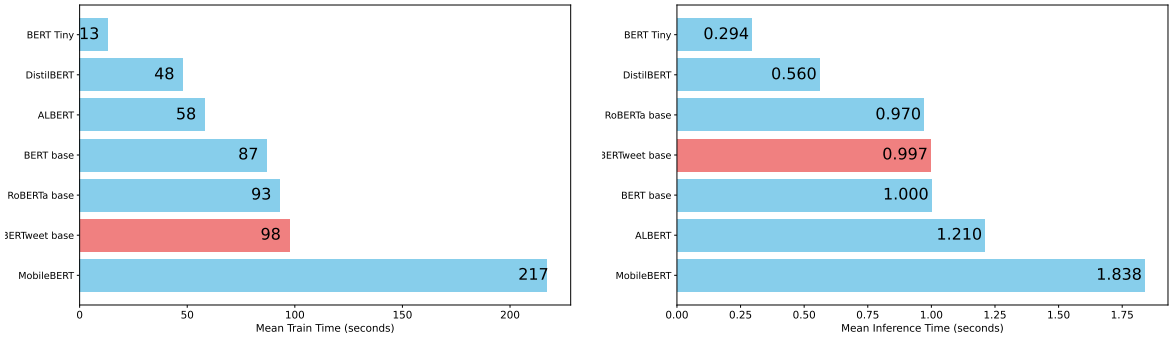
predictive capabilities [21].

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. 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 a lower train and inference times by 51% and 43% in comparison to BERTweet. For ALBERT the results are mixed. While the train time is lower than BERTweet by 40%, the inference time is much higher than all models except for MobileBERT.

4.2.3 Deep-dive into the results

TO UPDATE



(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.

4.3 Experiment set 2 results: Cross-domain results

TO UPDATE

Chapter 5

Conclusions

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu. In enim justo, rhoncus ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus. Aenean leo ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet. Etiam ultricies nisi vel augue. Curabitur ullamcorper ultricies nisi. Nam eget dui. Etiam rhoncus. Maecenas tempus, tellus eget condimentum rhoncus, sem quam semper libero, sit amet adipiscing sem neque sed ipsum. Nam quam nunc, blandit vel, luctus pulvinar, hendrerit id, lorem. Maecenas nec odio et ante tincidunt tempus. Donec vitae sapien ut libero venenatis faucibus. Nullam quis ante. Etiam sit amet orci eget eros faucibus tincidunt. Duis leo. Sed fringilla mauris sit amet nibh. Donec sodales sagittis magna. Sed consequat, leo eget bibendum sodales, augue velit cursus nunc.

Bibliography

[1] Artstein, R., and Poesio, M. Inter-Coder Agreement for Computational Linguistics. *Computational linguistics - Association for Computational Linguistics 34*, 4 (2008), 555–596. Place: One Rogers Street, Cambridge, MA 02142-1209, USA Publisher: MIT Press.

[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. Place: London Publisher: Routledge.

[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. Place: CHICAGO Publisher: Journal of Consumer Research.

[4] Bhargava, P., Drozd, A., and Rogers, A. Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics, Oct. 2021. arXiv:2110.01518 [cs].

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

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

[8] 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. Place: AMSTERDAM Publisher: Elsevier B.V.

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

[10] 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. Place: Hoboken Publisher: Elsevier Inc.

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

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

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

[14] 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.

[15] 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.

[16] Nguyen, D. Q., Vu, T., and Nguyen, A. T. BERTweet: A pre-trained language model for English Tweets, Oct. 2020. arXiv:2005.10200 [cs].

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

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

[19] Pérez, J. M., Giudici, J. C., and Luque, F. pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks, June 2021. arXiv:2106.09462 [cs].

[20] Rook, D. W., and Fisher, R. J. Normative Influences on Impulsive Buying Behavior. *The Journal of consumer research 22*, 3 (1995), 305–313. Place: CARY Publisher: University of Chicago Press.

[21] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, Feb. 2020. arXiv:1910.01108 [cs].

[22] 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. Place: Helensburg Publisher: Taylor & Francis.

[23] 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. Publisher: Taylor & Francis Group.

[24] Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., and Zhou, D. MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices, Apr. 2020. arXiv:2004.02984 [cs].

[25] Tripp, T. M., and Gregoire, Y. When unhappy customers strike back on the Internet. *MIT Sloan management review 52*, 3 (2011), 37. Place: Cambridge Publisher: Sloan Management Review.

[26] 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. arXiv:1908.08962 [cs].

[27] 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. arXiv:1706.03762 [cs].

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

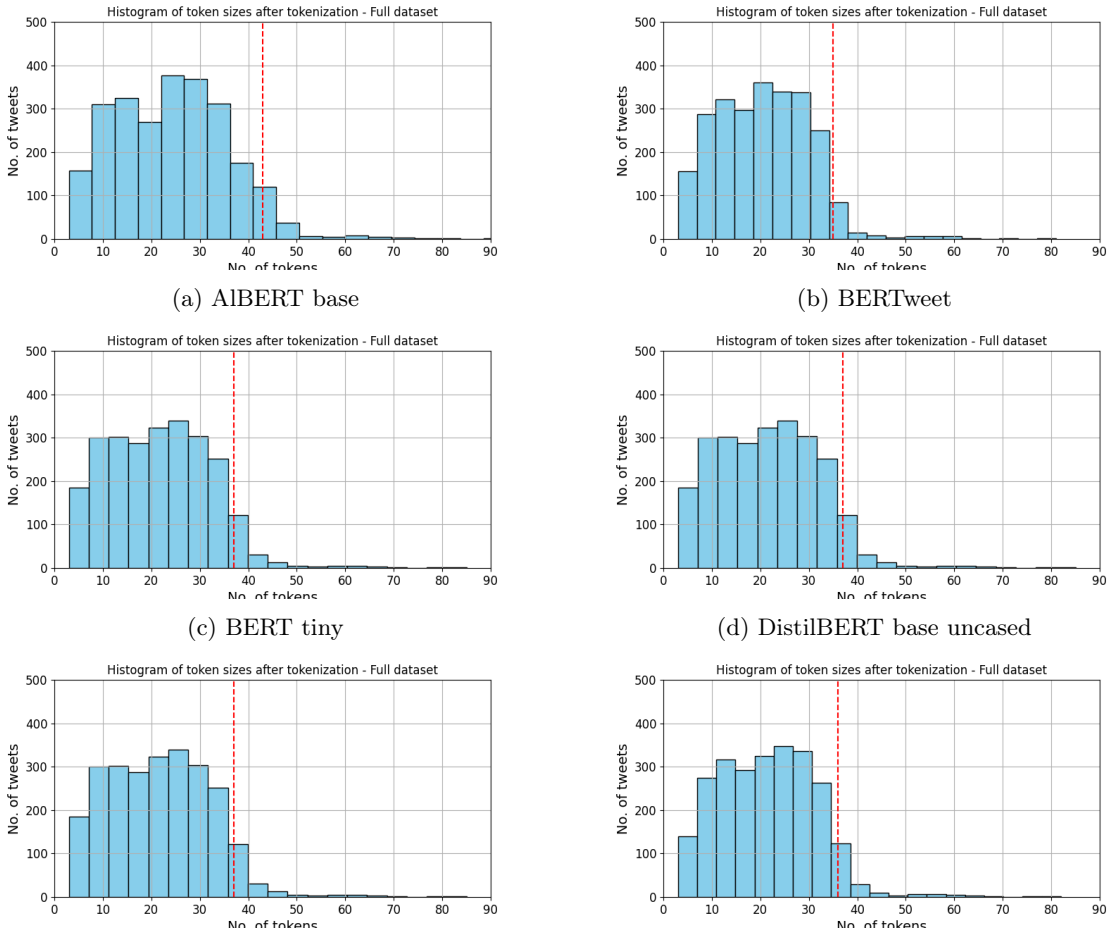
id	text	binarylabel	multilabel	domain
1.20E+17	asus g60 series . bought it to play games but guess not bf3	1	1	electronics
4.88E+17	love this trimmer ! making the sidewalk look sharp <url>	0	0	other

Table A.2: Sample data from [13].

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 3.1.



(a) AlBERT base

(b) BERTweet

(c) BERT tiny

(d) DistilBERT base uncased

(e) MobileBERT uncased

(f) RoBERTa base

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

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.