Exposing the Stance of Reddit Users Towards Brexit

Andrew Law

A thesis submitted for the degree of
Bachelor of Advanced Computing at
The Australian National University

November 2021

Except where otherwise indicated, this thesis is my own original work.

Andrew Law
15 November 2021

I want to dedicate this thesis to my mates. I wouldn't have been able to do it without you guys.

Acknowledgments

This work would not be possible without the help I've received from many people along the way. I would like to thank my supervisor Dr. Marian-Andrei Rizoiu for providing me with the opportunity to undertake this project and for his continuous guidance. I want to thank Duy Khuu for his collaboration and support on this project. I want to thank all the members of the Behavioural Data Science group for the support and community they have provided in this journey and in particular Rohit, Frankie and Tom for the assistance they have provided. I would also like to thank Jacob Pjetursson for providing feedback and advice on the project.

Abstract

Recently, social media has been blamed for the increasingly polarised nature of political discourse in our society. The ability to measure and combat political polarisation on social media is of significant importance if we wish to prevent polarisation from degrading the functioning of democracy and social cohesion. Stance detection provides a viable solution for addressing this problem, however so far no research has tested this technique on highly structured online discussions such as those found on the Reddit social media platform.

In this thesis, we propose a pipeline for annotating Reddit submissions for stance via crowdsourced workers from Amazon Mechanical Turk (MTurk). We conduct experiments to determine the optimum approach and parameters for conducting stance labelling with MTurk and produce a dataset of 5895 labelled r/Brexit submissions. We analyse the dynamics around r/Brexit discussions relating user activity to the occurrence of political events. We evaluate various novel strategies for improving BERT model performance on stance detection. Finally, we implement a state of the art stance detection model for Reddit user stance towards Brexit that achieves an F1 Score of 0.5547 compared to 0.3203 obtained by our previous baseline model.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The last decade has seen a meteoric rise in the popularity of social media as the world's preferred method of communication. Social media has made it easier than ever before for individuals to interact with each other and discover information provided by the wide user base of these platforms. At any given point in time, users are able to participate in countless discussions online through these platforms pertaining to a wide variety of topics ranging from the inconsequential to the controversial. Politics in particular, is a topic that is often the subject of fierce debate online and has consequently seen significant interest from researchers wanting to understand how social media influences the broader political landscape and democracy at large. Early pundits considered the open exchange of ideas on the internet as potentially yielding a more diverse and improved forum for political deliberation. Conversely, others have argued for the reverse effect as users may engage in homophily, which is the tendency for individuals to associate with similar individuals reducing the quality of political deliberation [Wojcieszak and Mutz, 2009].

More recently, it seems online political discourse has taken a turn for the worst. Try recall the last time you saw members from opposing sides of the political spectrum agree upon the same issue. Many would find this an increasingly rare occurrence in recent years. Social media has been blamed as the main culprit behind the rise in political polarisation within our society [Conover et al., 2011]. Although such platforms have been instrumental in connecting the world, the ubiquity and accessibility of social media has provided an alluring place for individuals to voice extreme and controversial views which may influence would-be moderates to more polarised viewpoints [Turner and Smaldino, 2018]. Social media has also been blamed for fostering "echo chambers" situations where users seek out other similar users to form closed groups in which beliefs are circulated and reinforced leading to polarisation [Bruns, 2017]. Additionally, the algorithms behind social media platforms have also been accused of exacerbating polarisation by producing "filter bubbles" the state of intellectual isolation that occurs from personalized search and information filtering preventing users from seeing alternate viewpoints [Bozdag, 2013].

1.1 Motivation

Political polarisation occurs when subsets of a population adopt increasingly dissimilar attitudes and positions towards political parties, policies and ideologies [Heltzel and Laurin, 2020]. To a degree, political partisanship is a necessary aspect of a well functioning democracy as it motivates individuals to participate in politics, engage in political debate and encourages diversity in policy alternatives [McCoy et al., 2018]. However severe polarisation may have far reaching negative consequences for democracy and society at large [Finkel et al., 2020].

With severe polarisation, groups increasingly antagonize and oppose each other, rendering productive debate and forming consensus on mutually beneficial policy impossible. At it's worst, polarised groups begin to view each other as existential threats and enemies to be fought and vanquished. This threatens to reduce social cohesion and generate turmoil and violence amongst the populace. In a highly polarized society, politicians are incentivised to adopt belligerent and anti-democratic tactics to rile their supporters which further exacerbates the deterioration of democracy and social cohesion. One only has to take a brief look at recent United States history to see stark examples of the potentially disastrous effects of political polarisation. From the rise of Donald Trump and the highly polarized nature of his election, to recent political unrest unfolding in the United States which saw liberals and conservatives battling each other on the street, to the ongoing debate around the COVID-19 vaccine and it's dangerous public health implications. It is without doubt that the world is awash with polarization and it's consequences.

As social media grows and the threat from online political polarisation continues to rise, researchers are increasingly looking for solutions to solve or mitigate these problems. Prior studies have found use in stance detection for combating the algorithmic issues of political polarisation on social media [AlDayel and Magdy, 2021]. By automatically knowing the stance of a user it is possible for algorithms to modify what kind of content they present to the user to avoid the polarising effects of filter bubbles, echo chambers and homophily. Stance detection can also provide many useful applications for studies analyzing public opinion. This might be applied in measuring the level of political polarisation within an individual or community or studying user pathways towards polarisation and the impacts of related phenomena in social media platforms. Successful applications of stance detection could see a reduction in political polarization online which could improve social cohesion, online political discourse and the effectiveness of democratic processes.

1.2 Research Questions and Contributions

So far most of the research conducted on stance detection and social media political polarization has been done within the context of the United States and the Twitter platform [AlDayel and Magdy, 2021]. Crucially, these studies have yet to capture the dynamics of stance and polarisation as they occur in more structured online discussion formats. To address this gap in knowledge, we focus this work on building

stance detection for the online discussions around the highly polarised Brexit debate.

We address several open research questions which have yet to be answered in the literature. The first question concerns gathering ground truth stance data on Brexit discussions which is a prerequisite for training supervised machine learning models required to achieve state of the art performance in stance detection. Therefore, we ask **how can we empirically quantify the stance of users around Brexit discussions?** To address this we gather discussion data from Reddit, a structured discussion based social media platform and propose a pipeline for annotating Reddit data via the Amazon Mechanical Turk crowdsourcing platform. We empirically determine optimal annotation pipeline parameters and experiment with different approaches for selecting workers to produce high quality labelled data. With this pipeline we obtain 5895 labelled texts for training our models.

Exploring the data we collect on Brexit Reddit discussions could yield valuable insights into the dynamics around online Brexit discussions and help inform the construction of stance detection models. Therefore in our second question we ask **what were the dynamics around Brexit discussions and how did they relate to major real world events?**. To address this we conduct a longitudinal and event based profiling of the data.

Finally, to address the main goal of this work, we ask the question **how can we build an accurate stance prediction around Brexit discussions?**. To this end we primarily leverage the BERT transformer based language model. We evaluate several novel strategies for improving BERT stance detection performance, namely, BERTweet, in-domain pre-training and multi-task fine-tuning with generated datasets. We achieve an F1 score of 0.5547 with our best performing model compared to 0.3203 obtained by the baseline, a weakly supervised, Twitter transfer learning Naive Bayes model.

In summary our main contributions will be:

- Pipeline for annotating r/Brexit submissions for stance via Amazon Mechanical Turk
- Dataset of r/Brexit submissions labelled for stance.
- Analysis of the dynamics of Reddit discussions around Brexit.
- State of the art stance detection model for classifying stance towards Brexit in online discussions.

1.3 Thesis Outline

In this thesis I will present my work over the course of six chapters. In Chapter 2, I will cover the background and introduce the literature surrounding the state of the art in stance detection. In Chapter 3, I will describe the process through which we source a dataset and generate labels. In Chapter 4, I will describe the methodology

we use to build a stance detection model. In Chapter 5, I will present the results of our experiments in stance detection. Finally in Chapter 6 I will summarise the work and discuss ways of extending the research.

Chapter 2

Background and Related Work

The following chapter introduces the literature and background knowledge required to understand our approach and methodology. In section 2.1 we will summarise the state of the art in stance detection. In section 2.2 we will discuss prior work done around information diffusion in online communities which this research in part builds off. In section 2.3 we will summarise the literature on the BERT transformer model. Finally in section 2.4 we will provide some background knowledge about Reddit and discuss related literature.

2.1 Stance Detection

Stance detection on social media is concerned with the automatic classification of an individual's stance towards a subject through information related to their posts and online activity [Küçük and Can, 2020]. Stance detection is often defined in different ways depending on the context and application at hand. We define stance as the overt expression of the speaker's thoughts, attitude and judgement towards a given proposition.

Most early approaches to stance detection relied solely on content features with traditional machine learning algorithms to model stance. These features infer stance by identifying similar vocabulary and linguistic features amongst individuals of a certain stance. SemEval 2016 Task 6A utilized a SVM classifier trained on character and word N-grams to produce a baseline model which achieved an average F1 score of 68.98, outperforming all submissions, including more sophisticated deep learning approaches [Mohammad et al., 2016]. More recently, researchers have begun to utilize network features in conjunction with content features to train stance detection models. Motivated by homophily, this approach aims to capture similarities in user behaviours such as likes, retweets, follows etc. to infer stance. Aldayel and Magdy [2019] produce a model for SemEval 2016 Task 6A using an SVM model trained using character and word N-grams along with features which model user interactions, preferences and connections over twitter, achieving an average F1 score of 72.49.

Although deep learning approaches initially failed to provide a performance boost over traditional machine learning models, the emergence of transformer based models has led to large performance improvements across many NLP problems, in-

cluding stance detection. Ghosh et al. [2019] fine tune a pretrained BERT model with raw Tweet text label pairs, achieving an average F1 score of 0.751. BERT not only achieves state of the art performance with no additional task specific training or tweaking, but is likely to generalize well across stance detection tasks where platform specific network features are not always be available. Hence we aim to use BERT in building our stance detection model.

2.2 Prior Work

The model we aim to implement was largely motivated by, and in part builds off of research conducted earlier by Mardale [2019] exploring information diffusion in online communities. This work primarily aims to understand how information exposure in online social networks effects people's opinions towards contentious issues. This is investigated through analyzing discussions occurring on the r/Brexit subreddit between November 2015 and April 2019 and developing a model for predicting the future stance of Reddit users towards Brexit.

This model uses the stance of the user at the current time and features related to user activity and the structure of diffusions the user participates in to predict the user's future stance. However as the author does not possess the necessary ground truth labels for Reddit user stance towards Brexit, they instead rely on a weakly supervised transfer learning approach using a Naive Bayes classifier trained on Twitter data to obtain their stance labels for Reddit. The Twitter training data is labelled for pro and anti Brexit stance based on the presence of certain hashtags chosen by the author to denote pro or anti stance positions. These model limitations cast significant doubt on the accuracy of any stance labels derived from this method. In section 5.1, we find that this approach yields an F1 Score of 0.3303, performing worse than random chance. We seek to address this with implementing a more robust model in our research.

2.3 Bidirectional Encoder Representations from Transformers (BERT)

Bidirectional Encoder Representations from Transformers, otherwise known as BERT is a new language representation model which produces state of the art performance on a wide variety of NLP tasks [Devlin et al., 2018]. BERT uses transformer architecture which utilizes the attention mechanism, a technique for dynamically weighting the importance of each element of an input sequence depending on the context [Vaswani et al., 2017]. This approach allows for modelling of global dependencies between input and output irrespective of distance between input elements. Unlike prior deep learning approaches which handle sequential data in order, transformers possess the context for any position in the input sequence, hence they are not limited to reading text in one direction and draw a greater understanding from input. This also has the added benefit of allowing significantly more parallelization.

With these capabilities afforded by the transformer architecture, BERT is pre-trained with two unsupervised machine learning tasks, Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). Pre-training is conducted on BookCorpus (800M words) and text passages from English Wikipedia (2.5B words) enabling the model to learn universal language representations.

Masked Language Modelling involves feeding the model input sequences with a portion of words randomly masked and training the model to predict the original value of the masked words based on the context available. This enables the model to learn bidirectionally and capture the meaning of words as they appear in different contexts as opposed to traditional word embedding approaches which assign fixed values to the meaning of words.

In Next Sentence Prediction, the model is fed labelled sentence pairs and tasked with predicting whether the second sentence follows the first sentence in the original document. During training, the labelled sentence pairs are generated from a corpus with half of the pairs containing adjacent sentences and half containing random sentences. This training is done so that the model has an understanding of the relationship between sentences which cannot be directly captured by MLM. This capability is an important part of many NLP tasks such as Question Answering and Natural Language Inference.

Applying BERT to downstream tasks is a straightforward process which involves initialising the BERT model with the pre-trained parametesr, plugging in task specific inputs and outputs into the model and fine-tuning all parameters end to end. In the case of text classification tasks like stance detection, the final hidden state of the classification token [CLS] which is at the start of every sequence is used to represent the sequence. Then a classification head is added to the BERT model to predict the class labels.

2.4 Reddit

Reddit is a relatively new social media platform that has seen significant growth in recent years, with over 18% of US adults having reported using the site in 2021 Auxier and Anderson [2021]. However so far there have been few studies on stance detection conducted in relation to Reddit which we aim to address with this research.

Reddit is unique amongst the mainstream social media platforms as it is designed for facilitating discussions in a highly structured manner as opposed to Twitter or Facebook which are designed more for broadcasting information and social networking. In addition Reddit users are almost always anonymous and there is no method for verifying one's identity on the platform unlike Twitter or Facebook. Reddit groups discussions around subreddits, forums where users are encouraged to post and comment about a specific topic or community. Subreddits are denoted with the prefix r/, followed by a user defined name e.g. the Brexit subreddit is denoted r/Brexit. A new discussion thread begins with a user submitting a post which typically contains a question, announcement, news or link to some media etc. to the

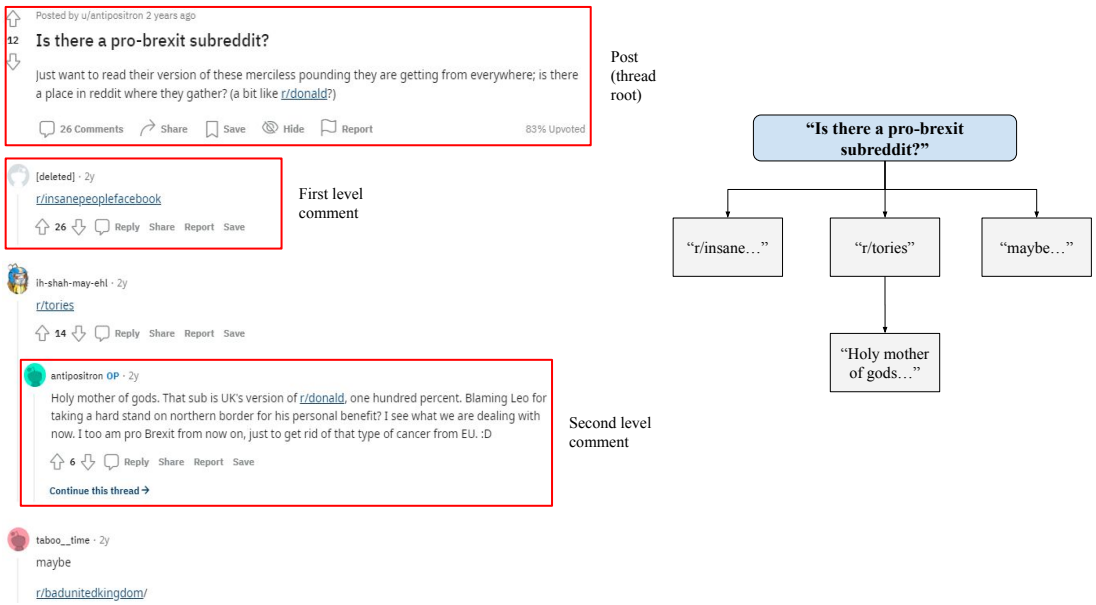

Figure 2.1: a) Post taken from the r/Brexit subreddit showing structure and features. b) Diagram illustrating tree structure of Reddit post shown on the left.

Subreddit. Other users are then able to comment replies to the post. Each comment within a thread can also have it's own comments, thus the discussion plays out in a tree like structure with the tree root at the post, as is shown in Figure 2.1. User's are also able to anonymously upvote or downvote comments or posts which enables a form of community self regulation. Popular submissions tend to be shown with higher priority to user's whereas negatively scored submissions are buried.

The result of these unique characteristics is that user's tend to have more agency for choosing the type of content they view by participating in particular subreddits. In doing this users may engage in homophily and form echo chambers within these subreddits. This may explain the commonly observed phenomenom where different Subreddits tend to have either a left or right leaning political bias [Soliman et al., 2019]. These characteristics may also have an effect on how user's express their stance towards contentious issues and how stance may be identified which we intend to investigate in this work.

Chapter 3

Annotating Brexit Stance

In this chapter we detail the approach we use for gathering labelled data for Reddit user stance towards Brexit. In section 3.1 we introduce the Reddit dataset and provide a longitudinal analysis of the data. In section 3.2 we give some background on crowdsourcing with Amazon Mechanical Turk. In section 3.3 we outline the design of our stance annotation task. In section 3.4 we describe how we ensure annotation quality and present our experiments for determining the optimal pipeline parameters and approach for selecting MTurk workers. Finally in section 3.5 we analyse the results of our stance annotation task.

3.1 Reddit Dataset

As there is no publicly available dataset for Reddit Brexit discussions, we gather our own dataset. We source Brexit discussion data from the r/Brexit subreddit which is the most popular Brexit related subreddit on the Reddit platform. We extend the r/Brexit dataset collected by Mardale [2019] by collecting additional r/Brexit submissions from May 2019 up to the end of February 2021. This data is collected from the Reddit Pushshift API using a simple Python webcrawler script [/r/datasets mod team, 2019]. The Reddit Pushshift API was created to provide enhanced search and access functionality for Reddit data. In addition the API acts as an archive for Reddit submissions enabling us to view submissions which have been deleted by moderators or the author, hence it is the preferred method for accessing Reddit data.

In total, our dataset spans from November 2015 to February 2021 and accounts for 815938 comments and 56017 posts, totalling 871955 submissions. In Figure 3.1 we present a time distribution of Reddit posts and comments in our dataset. From this figure we observe that the post and comment activity increases significantly after 2019 and remains quite high compared to prior years. This is possibly explained by increased growth of the Reddit platform in recent years and increased attention towards the Brexit debate. We also note significant activity surges in 29/03/2019, 09/09/2019 and 31/12/2020 which coincide with the occurrence of high profile events during the Brexit saga. This suggests that Reddit discussion activity around Brexit is highly volatile and dependent on the occurrence of events to drive discussion.

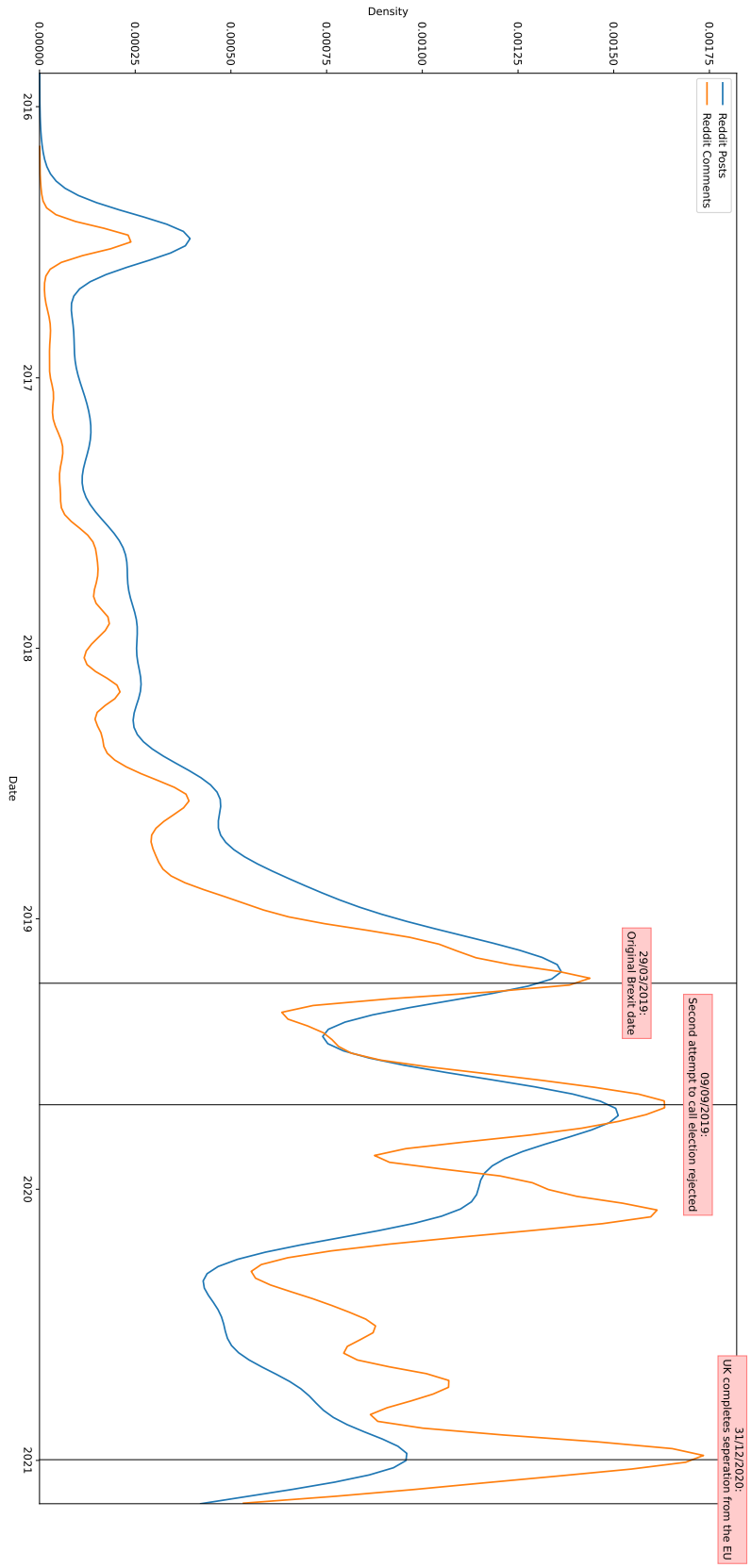

Figure 3.1: Time distribution of comments and posts collected from the r/Brexit subreddit between November 2015 and February 2021

3.2 Crowdsourcing

State of the art supervised machine learning models such as BERT require large quantities of labelled data to be trained effectively, however collecting this data has traditionally been expensive and time consuming. In recent years, crowdsourcing has become a popular option for researchers to gather large quantities of data due to it's low cost, high speed, flexibility, ease of use and diverse available worker pool [Callison-Burch and Dredze, 2010]. Crowdsourcing is particularly well suited for gathering labelled data as these tasks tend to be very simple, requiring little to no prior training or knowledge to execute. Therefore we employ the Amazon Mechanical Turk (MTurk) crowdsourcing platform to gather ground truth stance labels for our r/Brexit data [Amazon Mechanical Turk].

The MTurk platform enables requesters to submit human intelligence tasks, otherwise known as HITs to the global available work pool. Requesters first specify their task and task parameters which include, number of unique annotations per HIT, payment for each HIT and optional qualifications which allow requesters to restrict the workers who can execute the task based on chosen conditions. Simple HITs can be created using presets and widgets provided by MTurk with more complex HITs created using bespoke HTML and Javascript. Requesters then submit batches to their task, which include the data required for each HIT. Crowdsourced workers, referred to on the platform as Turkers are then able to browse available HITs and choose which HITs they would like to complete. After HITs are completed requesters may approve or reject the completed tasks they have received, determining whether the worker is paid or not. Requesters can specify HITs, submit task batches and handle data programmatically using the MTurk API.

Despite the many benefits of using crowdsourcing for data annotation, there are significant limitations associated with this approach. Crowdsourced workers are non-professional and thus may not always competently fulfil requests. As hourly pay is determined by how fast workers can complete HITs, there is a strong incentive for workers to cut corners and rush work which may reduce data quality. Anonymity of workers on the platform also means they cannot be held accountable should they fraudulently complete HITs. In addition, the authenticity of workers cannot be verified easily. This has led to concerns over the prevalence of bots or bad actors within the MTurk worker pool and the potential negative impacts they might have on studies using the service [Chmielewski and Kucker, 2020]. Therefore it is critical that we design our annotation task to mitigate these limitations to ensure that our labels are of high quality and funds are used effectively.

3.3 Annotation Task Design

First we define the classes in our stance detection task. Following the approach taken by Mohammad et al. [2016] we opt for classifying stance into three classes, namely, pro-Brexit, anti-Brexit or neither. As we intend to use crowd workers to gather stance labels, it is important that our stance classification task is formulated in a way which

is easy for annotators to approach and hence we do not include a neutral stance or additional degrees of pro-Brexit or anti-Brexit. Prior empirical studies have found that annotating for neutral stance is often quite difficult to achieve. In polarized settings it is expected that neutral users will only make up a small portion of the population, such users do not tend to be explicit about their neutral position and the absence of favor or against signals cannot be used to infer neutrality either.

We design a simple questionnaire style MTurk interface which contains 5 questions per HIT for workers to annotate Reddit texts. The interface is implemented using HTML and Javascript and is partially shown in Figure 3.2. Additional detailed instructions (not shown) are provided to workers including directions for annotating stance and definitions for pro-Brexit, anti-Brexit and neither stances. We design our task so that unbroken portions of discussion threads are shown to workers for annotation. This enables workers to simultaneously use reply context to help infer the stance of submissions and determine the stance of adjacent submissions in the thread. To construct the data for each HIT, we randomly sample comments and posts from the dataset, then using the parent of each comment, we build sets of 2, 3, 4 and 5 length discussion threads. We arrange these different length threads along with singular posts into a CSV file with 5 Reddit submissions per row. We then submit this file to MTurk for annotation by workers.

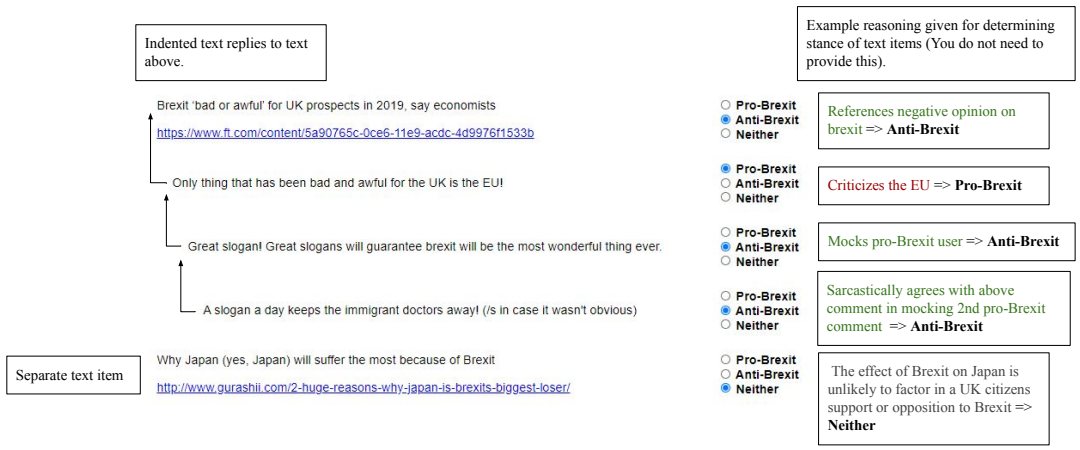

Figure 3.2: Instructions shown to workers in Amazon Mechanical Turk depicting an example HIT with tips for how the worker should interpret the interface and solve the task. Note the arrows and text boxes do not appear in the actual questionnaire.

To ensure the quality of our annotations, each HIT is annotated by 8 unique workers [Mohammad et al., 2016; Snow et al., 2008]. We measure the quality of annotations by computing the inter-annotator agreement (IAA) which is the proportion of annotations that agree with the majority stance label. Annotations which did not have an IAA >= 0.6 (i.e. atleast 5 out of 8 annotators agreeing on the majority label) were dropped from the final dataset. Reported IAA values are computed after

dropping unless specified otherwise.

To determine an appropriate payment for workers, we manually conduct annotation on a sample of Reddit comments and measure the average time taken to annotate 5 texts (i.e. a single HIT) to be 45 seconds. We extrapolate from the US federal minimum wage of $7.25 per hour to come up with a reward of $0.10 per HIT which equates to a rate of approximately $8 per hour. MTurk charges a flat 20% fee on top of rewards given to workers, therefore with 8 annotators we have a net cost of $0.96 per labelled text. To ensure we are in good standing with the MTurk community and our payment is consistent with what workers think is fair, we monitor Turkopticon and Turkerview throughout the annotation process [tur; TurkerView, 2021]. These online platforms enable workers to post anonymous reviews about requesters and the HITs they encounter to help other workers avoid working with unscrupulous requesters, identify worthwhile HITs and provide feedback to requesters. We received several reviews on Turkerview and find that our payment is consistent with what workers think is fair.

We set the HIT expiry to 7 days after posting and initially set the HIT completion time limit to 2 minutes. Later on we received feedback from a worker who asked for more time to complete our HITs so we increase the time limit to 10 minutes. Prior studies suggest that a dataset of approximately 4000 text instances is sufficient for training and evaluating classifiers for stance detection tasks [AlDayel and Magdy, 2021]. Therefore we aim to collect an initial dataset of 4000 text instances at a minimum and add more annotations as needed during the model training process.

3.4 Annotation Parameter and Context Tuning

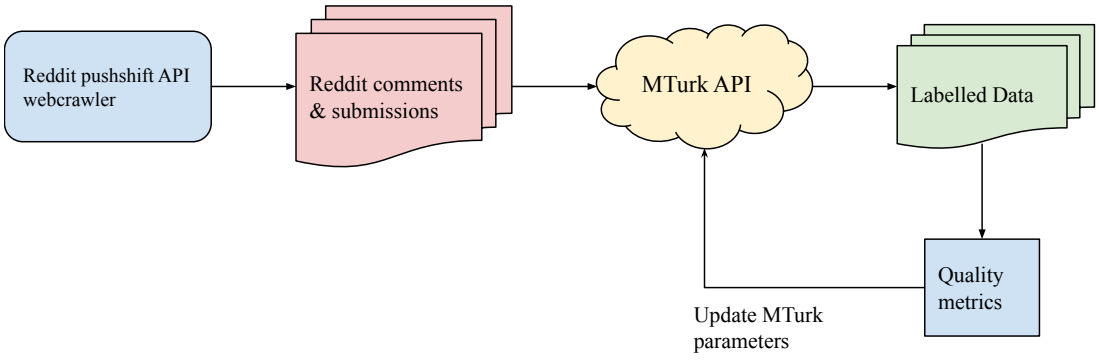

Figure 3.3: Data pipeline for producing Brexit stance labels.

Before going into full scale production, we first submit a series of small batches for annotation, each containing 200 text instances and different task parameters. We evaluate the IAA of these batches to identify optimal task parameters for our final setup. During this process, our initial starting point and adjustments we make to the

task parameters were informed by literature around best practices in MTurk [Ahler et al., 2019; Amazon Web Services, 2021]. Our general data pipeline is shown in Figure 3.3.

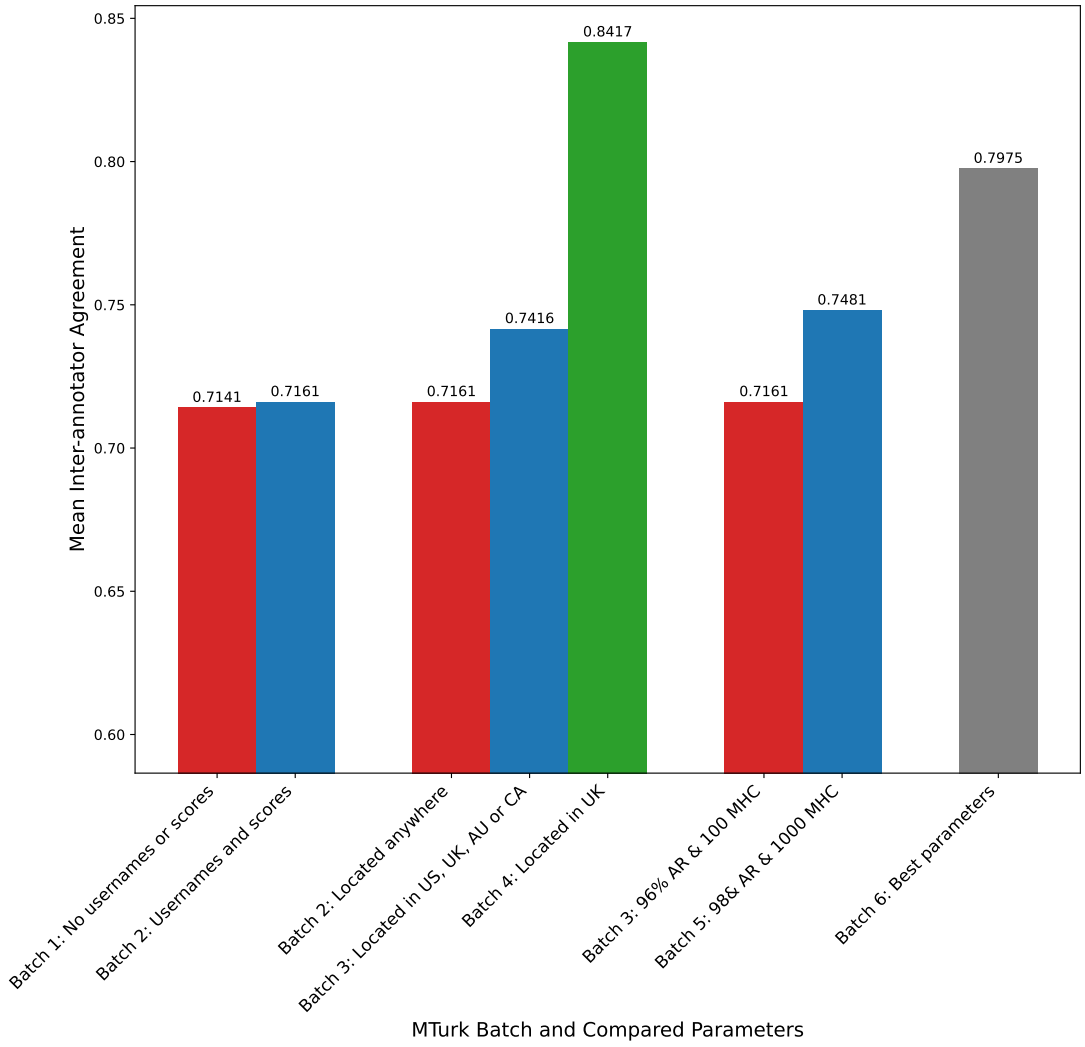

Figure 3.4: **MTurk parameter tuning:** each bar group represents a set of batches showing the effect on IAA of adjusting a particular parameter alone in MTurk

Table 3.1 shows the different parameters we experimented with and the associated IAA metrics for each preliminary batch. Figure 3.4 shows a grouped bar chart comparing the mean IAA for different options we experimented with and the performance of the best parameter set. First we experimented with showing usernames and scores associated with each Reddit submission. This is motivated by prior research which suggests that additional context can improve annotation quality or may hinder it if the context signals outweigh other signals [Joseph et al., 2017]. However this adjustment only resulted in a very minor increase in IAA from 0.7141 to 0.7161.

Batch Number	Minimum Approval Rate	Minimum HITs completed	Other Conditions	Context	Estimated Turnaround Time	Inter Annotator Agreement statistics (IAA) Count	Mean	Std	IAA after dropping agreement < 5/8 Count	Mean	Std	Data loss rate
1	96%	100	N/A	Discussion Threads	<1 day	200	0.604	0.152	105	0.714	0.117	0.475
2	96%	100	N/A	Discussion Threads, Usernames and Scores	<1 day	200	0.614	0.150	118	0.716	0.103	0.410
3	96%	100	Location in US, UK, AU or CA	Discussion Threads, Usernames and Scores	<1 day	200	0.674	0.145	149	0.742	0.097	0.255
4	96%	100	Location in UK	Discussion Threads, Usernames and Scores	Incomplete	200	0.693	0.187	35	0.842	0.127	N/A
5	98%	1000	N/A	Discussion Threads, Usernames and Scores	4 days	200	0.651	0.167	130	0.748	0.124	0.350
6	98%	1000	Location in US, UK, AU or CA	Discussion Threads, Usernames and Scores	4 days	200	0.740	0.172	163	0.798	0.135	0.185
7	98%	1000	Location in US, UK, AU or CA	Discussion Threads	4 days	200	0.720	0.174	154	0.789	0.135	0.230
8	97%	500	US, UK, AU or CA Qualification Test >= 80	Discussion Threads, Usernames and Scores	Incomplete	0	0	0	0	0	0	N/A
9	96%	100	US, UK, AU or CA Qualification Test >= 80	Discussion Threads, Usernames and Scores	Incomplete	0	0	0	0	0	0	N/A
10	98%	1000	Location in US, UK, AU or CA	Discussion Threads, Usernames and Scores	3 days	1000	0.671	0.156	725	0.746	0.110	0.275
10`	98%	1000	Location in US, UK, AU or CA **Bad worker entries removed**	Discussion Threads, Usernames and Scores	N/A	1000	0.757	0.111	657	0.866	0.111	0.343
11	98%	1000	Location in US, UK, AU or CA	Discussion Threads, Usernames and Scores	14 hrs	1000	0.597	0.145	543	0.710	0.092	0.567
11`	98%	1000	Location in US, UK, AU or CA **Bad worker entries removed**	Discussion Threads, Usernames and Scores	N/A	1000	0.716	0.189	425	0.908	0.083	0.675
12			Cherry picked workers	Discussion Threads, Usernames and Scores	3 days	200	0.796	0.182	176	0.840	0.146	0.120

Table 3.1: MTurk parameter tuning batch data. Prime batches (e.g. 10′) represent the batch metrics recomputed after we remove annotations produced by bad workers.

We saw a significant increase in IAA when restricting workers to US, UK, AU or CA locations from 0.7161 to 0.7416 with minimal effect on batch turnaround time. When limiting workers to UK locations we saw an even greater increase in IAA up to 0.8417. This is fairly intuitive given the subject of our annotation is Brexit. Workers living in the UK are likely to be more familiar with the jargon and terms used in the Brexit debate and would have an easier time inferring stance. However, this batch was only partially complete after 7 days and thus was unfeasible for use in future batches. Experimenting with increasing approval rate (AR) from 96% to 98% and minimum HITs completed (MHC) from 100 to 1000, we saw a significant increase in IAA from 0.7161 to 0.7481 with an increase in batch turnaround time from <1 day to 4 days.

After conducting these experiments we found showing usernames and scores, Restricting workers to US, UK, AU or CA locations, 98% AR and 1000 MHC to be the optimal parameter setup for our annotation task. These parameters yielded an IAA of 0.7975 which is close to our desired IAA baseline established by Mohammad et al. [2016] of 0.8185 and a turnaround time of approximately 4 days.

3.4.1 Qualification Testing

In addition to these task adjustments, we experimented with employing a Qualification Test to screen workers for competency in annotating r/Brexit texts for stance [Callison-Burch and Dredze, 2010]. The test is unpaid and consisted of a questionnaire with 10 handpicked r/Brexit texts which are annotated for stance similar to the format in Figure 3.2. Workers receive 10 points for each right answer and those who achieve a score above a threshold are granted permission to complete the HITs.

In batch 8 we employ the Qualification Test, qualifying workers who score at least 8/10 correct, have 97% AR and 500 MHC, however we received no annotations after the HIT expired. In batch 9 we loosen AR to 96% and MHC to 100 to provide a larger worker pool for our HIT, however we only receive 20 complete annotations out of 1600 at HIT expiry. Further analysis of testing results showed that in batch 9, 47 workers completed the Qualification Test, 11 workers scored >= 80 and 31 workers scored >= 60.

Demonstrably, such an approach is ineffective for generating sufficient worker volume for our HITs to be completed on time. There are several reasons for why this might be the case. When the worker submits the qualification test they are re-directed back to the main dashboard. The worker is thus required to search for the HIT again which may not be visible on the dashboard anymore. If the test is set to auto-grade and auto-grant the qualification, there is a delay of several minutes, so even after completing the test workers cannot immediately access the HIT. These reasons may contribute to workers loosing interest in completing our HIT after successfully clearing the Qualification Test. We also potentially miss a large portion of competent workers who choose to filter out HITs which they are not qualified for or ignore Qualification Test HITs. Although these issues could be mitigated by paying workers to complete a Qualification Test via a HIT, we did not deem this an efficient use of

funds.

3.4.2 Suspicious Worker Behaviour

Despite employing optimal parameters, our first production batch of 1000 instances (batch 10) yielded an IAA of 0.746, significantly below the expected IAA of 0.798. The next production batch (batch 11) fared even worse with an IAA of 0.710. We manually inspected the annotations in these batches and found a pattern of suspicious behaviour exhibited by certain workers suggesting the presence of bots or bad actors. The data we collected up to this point suggested the majority of stances to be neither with a small number of anti-Brexit stances and even fewer pro-Brexit stances expressed. However these workers produced annotations which were either random or contrary to proportion of each stance we expect to find in their annotations.

To quantify this, for each worker we compute the Majority Agreement Proportion (MAP), which is the proportion of a workers annotations which are in agreement with the majority label. In batch 10 and 11 we find several prolific workers with an abnormally low MAP below 0.25 and add these workers to a list of workers banned from completing any of our HITs. We then remove annotations produced by these workers from batch 10 and 11 and recompute IAA, yielding IAA scores of 0.866 and 0.908 respectively. To avoid further incidents of bots or bad actors, we scan all our existing annotations to create a white list of workers who have annotated at least 20 instances with a MAP > 0.5. In batch 12 we limit HIT eligibility to workers in this white list and we achieve an IAA of 0.840 with a turnaround time of approximately 3 days using this approach. Therefore, we continue with the production of the remaining Brexit annotations using this white list approach.

Although 98% AR with 1000 MHC in conjunction with location filtering seems like a fairly strict criteria for quality, in practice many requesters are reluctant to reject HITs potentially allowing bad actors to slip through. This occurs even if the HITs are obviously fraudulent for fear of reprisal on review sites which could damage a requester's reputation and cause them to be blacklisted by the MTurk community. As shown in our annotation results, the prevalence of bad workers may already by widespread. Existing worker quality controls using qualifications are insufficient to screen out bad workers, necessitating the use of manual vetting of workers to produce high quality stance annotations.

3.5 Stance Annotation Analysis

In total we annotate 7543 texts and produce a final dataset of 5895 labelled texts with an IAA of 0.804. In the final dataset 295 instances are labelled pro-Brexit, 4521 instances are labelled neither and 1079 instances are labelled anti-Brexit. This suggests that the r/Brexit subreddit is a left leaning community. Neither instances account for over 76% of the dataset which is considerably higher than the proportion of neither instances found in the SemEval 2016 Task 6A twitter political stance dataset [Mohammad et al., 2016]. This suggests that Reddit users are less likely to overtly express

their stance towards topics when engaging in discussions compared to their Twitter counterparts.

Class	IAA before dropping	IAA after dropping
Pro-Brexit	0.6228	0.7197
Neither	0.7585	0.8244
Anti-Brexit	0.6937	0.7417

Table 3.2: Mean inter annotator agreement between classes

Table 3.2 shows the mean inter annotator agreement between classes. Here we observe that annotators tend to have the most agreement when annotating neither instances and the least agreement when annotating Pro-Brexit instances. This phenomenon is most pronounced before dropping instances. It is possible that pro-Brexit texts exhibit language which workers are less familiar with, making these texts inherently more difficult to annotate. Conversely, the language of neither texts may contain less group specific jargon making them easier for workers to annotate. Additionally, if we treat HITs as a kind of training for detecting Brexit stance, workers may be less trained for the stances that they least encounter, resulting in more errors and lower IAA for infrequent stances.

Chapter 4

Stance Detection Methodology

Although BERT models achieve state of the art performance in many NLP tasks, current methods for fine-tuning BERT models are still quite unsophisticated. There is little research on enhancing BERT fine-tuning methods to improve performance on downstream tasks. In this chapter we introduce our methodology for fine-tuning BERT and discuss three strategies for improving BERT performance, namely, BERTweet, in-domain fine-tuning and multi-task fine-tuning. Our general framework for enhancing BERT performance is shown in Figure 4.1. We also discuss a method for improving overall stance detection performance through data augmentation.

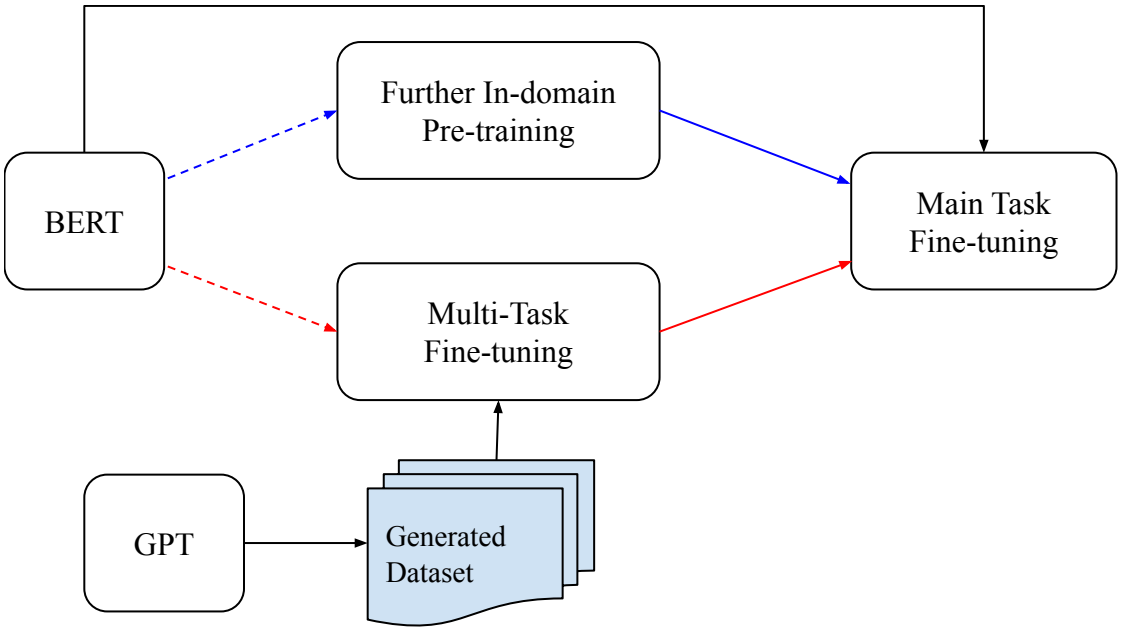

Figure 4.1: Framework for enhancing BERT performance which shows three methods for training BERT on a downstream task.

4.1 BERT fine-tuning

To perform our experiments, we utilise the Huggingface Transformers python library and Pytorch to implement our BERT models [Wolf et al., 2020; Paszke et al., 2019]. We use the default BERT-base model architecture which has 12 layers in the encoder stack, 12 attention heads and a hidden layer size of 768. Our minimum sequence length is 512 tokens and a softmax multi-class classification head is used to predict the stances. BERT fine-tuning is performed using 2 NVIDIA Quadro RTX 6000 GPUs.

We fine-tune BERT using a randomized hyperparameter tuning setup for 60 iterations with a train test split of 70%/15%/15% forming the training, validation and test sets respectively Gallicchio et al. [2017]. During each hyperparameter tuning iteration, hyperparameters are randomly chosen from a predefined search space and a BERT model is trained. During each epoch of training, the model is evaluated with the validation set and the best performing model is saved. The iteration which produces the model with the best validation score determines the optimal hyperparameters and is used for final test evaluation. The hyperparameter search space is shown in Table 4.1.

Hyperparameter	Search Space
Number of Epochs	{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15}
Batch Size	{16, 17, 18, 19, 20, 21, 22, 23, 24}
Learning Rate	(0.00001, 0.00006)

Table 4.1: Hyperparameter search space used in BERT model training. Curly brackets denote a discrete set of values and round brackets denote a continuous value range.

We experimented with using learning rate scheduler with warmup and linear decay in a 1:10 ratio of warm up steps to training steps however this failed to increase performance. We also dropped weight decay from our hyperparameter tuning setup as early experiments with it showed that it did not improve model performance.

4.2 BERTweet

Language on social media is often quite different from the language encountered in traditional corpora [Eisenstein, 2013]. Improper grammar, internet slang, abbreviations and spelling mistakes are all common occurrences on social media which are absent from traditional corpora. This poses a challenge for applying models trained on traditional corpora to social media text. In this work we experiment with two pre-trained BERT model variants, the original BERT-base (uncased) model and BERTweet (cased) developed by Nguyen et al. [2020].

BERTweet uses the same architecture as BERT-base, however it is pre-trained on Twitter data using the RoBERTa pre-training procedure [Liu et al., 2019]. This Twitter corpus is composed of 850M English Tweets which contain 845M Tweets sourced from between January 2012 and August 2019 and 5M Tweets related to COVID19.

Therefore, if the language of r/Brexit is more similar to Twitter language than that of Wikipedia and BookCorpus, we expect BERTweet to outperform BERT-base on our stance detection task.

4.3 In-domain Pre-training

As Nguyen et al. [2020] show in their BERTweet study, by aligning the pre-training corpus to the target domain we can better adapt BERT to the target task, improving performance. However, training BERT language models from scratch requires a massive amount of data and computational resources which most researchers either do not have access to, or cannot justify for a single downstream task. A natural compromise is to further pre-train an already trained BERT model with additional data from the task domain Sun et al. [2019].

To implement this strategy, we further pre-train BERT on the masked language modelling task with our corpus of 871955 r/Brexit submissions. We refer to this model as BERT-Reddit. To run the training process we utilize the MLM python script provided by Huggingface. Training of BERT-Reddit was performed using 8 NVIDIA V100 GPUs in a distributed fashion [National Computer Infrastructure, 2021]. The time consuming and costly nature of MLM pre-training makes exhaustive hyperparameter tuning impractical, therefore we experiment with a small set of hyperparameters which we optimise through trial and error. Pre-trained models are evaluated based on their performance on the downstream stance detection task. Throughout training we use a learning rate scheduler with warmup and linear decay in a 1:10 ratio of warmup steps to training steps. Eventually we settle on a batch size of 128, learning rate of 5.0e-05 and 6 epochs of training as our best hyperparameters.

4.4 Multi-Task Fine-tuning

Multi-Task fine-tuning is a strategy proposed by Sun et al. [2019] which involves exploiting the shared learning between different tasks in the same target domain to improve performance on the main task. Although the authors find Multi-Task fine-tuning to be less effective than in-domain pre-training we might be able to yield some benefit from this approach if used in conjunction with other BERT enhancement strategies.

4.4.1 Generating Text Similarity Datasets with DINO

To perform Multi-Task fine-tuning we first must have alternative tasks available to us in the domain of Brexit discussions. However we do not have access to any alternative Brexit discussion datasets that would provide us with different tasks to fine-tune on. Therefore we implement a novel approach proposed by Schick and Schütze [2021] for unsupervised fine-tuning of BERT models on semantic textual similarity (STS) using generated datasets referred to as Datasets from Instructions (DINO).

Task: Write two sentences that mean the same thing.
Sentence 1: "When he was elected mayor of London."
Sentence 2: "When he became mayor of London."

Task: Write two sentences that are on completely different topics.
Sentence 1: "When he was elected mayor of London."
Sentence 2: "He's just looking in the mirror."

Figure 4.2: Example of intended sentence output from DINO, given a sentence input and instruction. Similar and dissimilar sentences are labelled accordingly to form a Dataset [Schick and Schütze, 2021].

This approach leverages pre-trained language models (PLM) to artificially generate datasets of labelled text pairs which are used for fine-tuning. To generate labelled datasets, the PLM is fed instructions and a set of texts in the target domain. For each input text an output text is generated based on the instruction. These texts pairs are labelled according to their instruction to form a labelled dataset. An example of this is illustrated in Figure 4.2.

We undertake Multi-Task fine-tuning of BERT using the STS task with a dataset generated from DINO. We use the python implementation provided by Schick and Schütze [2021] to conduct our DINO experiments. Our dataset is generated with our r/Brexit text corpus and GPT2-XL. During training we found that the self-debiasing part of the DINO training algorithm was bottle necking training speed. Therefore we parallelise training by segmenting the input dataset into 12 parts and running DINO with each segment on a separate GPU compute node (1 NVIDIA V100 GPU). We give DINO instructions to generate similar text pairs labelled 1 and dissimilar text pairs labelled 0 as shown in Figure 4.2. We generate 5 entries per input and label and delete identical pairs. The default hyperparameters suggested by Schick and Schütze [2021] are used. From this we generate a dataset of over 4M labelled text similarity pairs.

4.4.2 Fine-tuning BERT on Classifying Text Similarity

We fine-tune BERT with the generated DINO dataset on text pair classification and produce a model which we refer to as BERT-DINO. BERT-DINO is trained with 8 NVIDIA V100 GPUs in a distributed fashion. We set maximum sequence length to 128 tokens due to memory constraints. Batch size is set to 256, learning rate to 2e-05 and number of epochs to 3. We split the DINO dataset 95%/5% for training and validation sets and achieve an accuracy of 0.80 on the validation set.

4.5 Stance Dataset Augmentation

In addition to these model enhancement strategies we also experiment with augmenting the Brexit stance dataset to overcome any performance issues that may arise from the class imbalance. We experiment with two augmentation strategies, oversampling and automatic text labelling. With oversampling, we randomly sample instances from our minority classes to duplicate in the training set so that each class is equally sized. However early experiments revealed this to be unsuccessful so we did not continue with this approach.

In the second approach we apply heuristics to our r/Brexit dataset to automatically generate weakly labelled data for our minority classes. During qualitative analysis of Brexit Reddit discussions, certain keywords were found to be strongly associated with either pro or anti Brexit speech. For example, "remoaner" was a common derogatory term used by pro-Brexit users to disparage others who were anti-Brexit. We additionally found "remainiac" in use by pro-Brexit users and "Brexshit" in use by anti-Brexit users. We use the presence of these terms to create weak labels for pro and anti Brexit texts. We also identify users which have been annotated as pro-Brexit and qualitatively analyse their other Reddit submissions to gauge their true stance. Users which are found to be strongly pro-Brexit are marked and weak pro-Brexit labels are produced for the rest of their submissions which have not yet been annotated. With this approach we produce an additional **2145** pro-Brexit and **622** anti-Brexit weakly labelled texts.

Chapter 5

Results

In this chapter, we present and discuss the results of our experiments. In section 5.1 we show how our models are evaluated. In section 5.2 we describe the experimental setup of our baselines and compare the results with the results of our BERT models. In section 5.3 we discuss the results of the various BERT Fine-tuning strategies outlined in chapter 4. In section 5.4 we conduct an external validation to evaluate our data augmentation. Lastly, in section 5.5 we discuss issues related to the performance of our models which have not yet been mentioned.

5.1 Evaluation

In multi-class classification, Macro F1 score is defined as the mean of each classes' F1 score. To evaluate our models, we use Macro F1 score as the primary performance metric and Accuracy as a secondary metric. Macro F1 score is chosen as our primary metric because our stance detection task is an imbalanced 3 class classification problem. We compute F1 scores against two dummy classifiers, one which predicts uniform random stances and one which predicts only the most frequent stance which yield scores of 0.26 and 0.29 respectively. **Therefore, we consider the random chance baseline for this task to be an F1 score of 0.33 and an Accuracy of 0.33**.

Note for the following model evaluations we use a subset of the full Brexit Stance dataset with 5006 labelled texts as additional annotations were carried out after running the experiments.

5.2 Comparison of BERT with Baselines

To establish an initial baseline we evaluate the prior Naive Bayes Brexit stance classifier by Mardale [2019] with our newly acquired ground truth labels, yielding an F1 score of 0.32 which is worse than random chance. Given that the classifier is trained on weakly labelled Tweets this is unsurprising. Twitter language around Brexit is likely too dissimilar from Reddit Brexit discussions for this kind of transfer learning framework to succeed. We propose new baseline classifiers constructed from traditional supervised machine learning approaches trained on our dataset. We employ

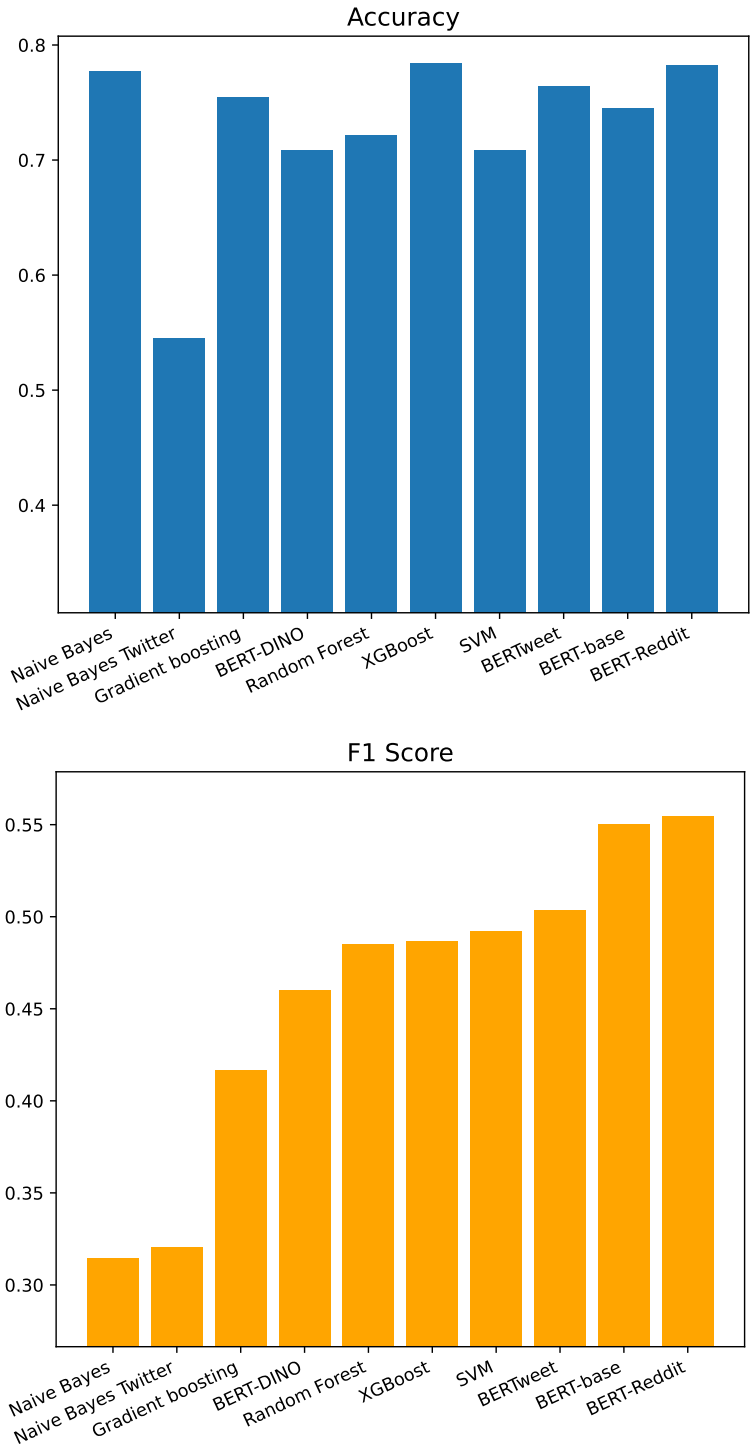

Figure 5.1: Visual comparison of Accuracy and F1 Score results for trained classifiers.

Classifier	Accuracy	F1 Score
Naive Bayes	0.7772	0.3145
Naive Bayes Twitter	0.5451	0.3203
Gradient Boosting	0.7546	0.4168
BERT-DINO	0.7084	0.4600
Random Forest	0.7215	0.4849
XGBoost	**0.7838**	0.4867
SVM	0.7082	0.4924
BERTweet	0.7639	0.5033
BERT-base	0.7454	0.5502
BERT-Reddit	0.7825	**0.5547**

Table 5.1: Results for trained classifiers.

Naive Bayes, Random Forest, SVM, Gradient Boosting and XGBoost classifiers in our baselines.

To train our baseline models we use TF-IDF features extracted from the stance dataset after stemming and dropping stopwords. Similiar to the setup used for training BERT, we use a randomized hyperparameter tuning setup to evaluate our baseline models and use the same train test split.

The results for our baseline models and BERT models are shown in Table 5.1 and in Figure 5.1 for visual comparison. We see that SVM performs the best amongst the baseline classifiers. This is consistent with prior literature where SVM tends to be commonly used for stance detection. Naive Bayes performs the worst with an F1 Score of 0.3145 which does not exceed the random chance baseline. Our best BERT model, BERT-Reddit achieves an F1 Score of 0.5547 which represents a moderate improvement over our baseline models.

5.3 Comparison of BERT Fine-tuning Strategies

The results show that BERT-Reddit yields a small performance improvement over the standard BERT implementation, whereas the BERTweet and BERT-DINO approaches degrade performance compared to BERT-base. Although the performance increase of BERT-Reddit compared to BERT-base is only slight, BERT-Reddit has a significantly higher accuracy than BERT-base which substantiates BERT-Reddit as an improvement over BERT-base. It is possible that the dataset we use of approximately 860000 texts for in-domain pre-training does not sufficiently adapt the model towards our task and more training data is needed to achieve more significant performance gains. Given the poor performance of BERTweet, there is evidence to suggest that the language used on Reddit is more similar to the language of traditional corpora such as Wikipedia and BookCorpus than to the language used on Twitter.

Somewhat surprisingly, the Multi-Task fine-tuning approach used in BERT-DINO resulted in a significant decrease in performance below many of our baseline clas-

sifiers. To investigate this issue, we analyzed some of the STS text pairs that were generated through DINO and uncovered a strew of low quality and incorrectly labelled pairs. If the proportion of these text pairs was large enough, fine-tuning with them could have caused the model to become significantly more biased. The dataset used here is also extremely large at approximately 5 million text pairs.

As training PLM models with large datasets can be very time and resource intensive we did not have the option to experiment with many different sets of hyperparameters and had to rely on mostly default settings. Lacking BERT performance in either the in-domain pre-training or multi-task fine-tuning strategies could be explained by improper assignment of hyperparameters, which could be causing the model to over fit, or insufficiently learn from the MLM or STS task, resulting in poor performance on the stance detection task. This could also be the case with our dataset generation in DINO if our the hyperparameters of our GPT2-XL model were not configured properly.

5.4 External Validation

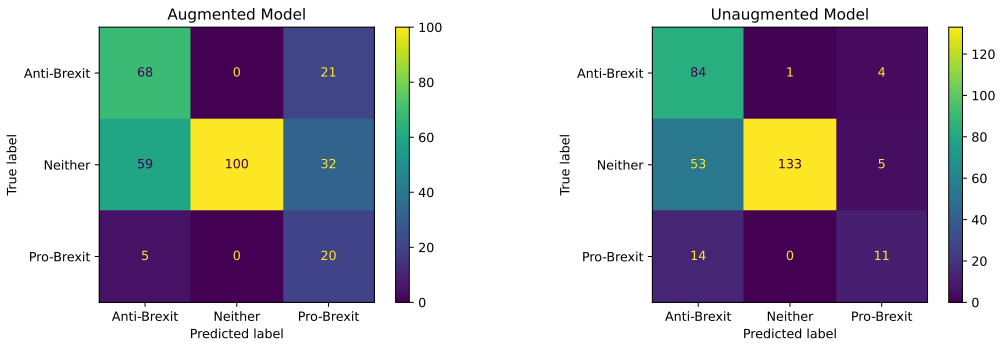

Figure 5.2: Confusion matrices for external validation with augmented (left) and unaugmented (right) models.

Training Data	Accuracy	F1 Score
Augmented	0.6164	0.5703
Unaugmented	0.7475	0.6691

Table 5.2: Results for external validation with augmented and unaugmented models.

To evaluate our data augmentation approach we employ an external validation method, similar to the active learning strategy proposed by Kong et al. [2021]. This involves training our model with the full dataset and selecting for 100 annotation instances which our model is most confident in predicting for each class. These annotations then constitute a new dataset which is used to evaluate the model. We use

this approach to test our BERT-Reddit trained with augmented and unaugmented datasets. Results are shown in Figure 5.2 and Table 5.2. Comparing the confusion matrices shown in Figure 5.2, we observe that the model fails to improve upon it's ability to predict the pro-Brexit, seemingly resorting to aggressively and somewhat randomly predicting the pro-Brexit stance. This is despite significant dataset augmentation with pro-Brexit labels. Table 5.2 shows that the unaugmented significantly outperforms the augmented model in both F1 Score and Accuracy.

We qualitatively analysed the automatically labelled texts to investigate the performance issues and noticed that several Reddit comments quoted pro-Brexit keywords and sometimes entire sections from comments made by other users. This poses an issue for our automatic stance labelling where anti-Brexit users may be quoting pro-Brexit keywords leading to erroneous augmented labels. In addition, the issue of users quoting other users poses a larger challenge for language models which are not able to discriminate between what is quoted and what isn't.

In this case, such a model would interpret the quote as part of the stance of the whole text resulting in incorrect predictions. This is perhaps a unique aspect of Reddit as users writing on Facebook or Twitter do not tend to quote others and include no special features for making quotations. Reddit in contrast, tends to facilitate much longer, more substantive and thought out debate which may involve more quoting. Reddit also includes features on their platform to make it very easy to quote others and format quotations in a pretty way, encouraging users to quote things. Even if it is not necessary, Reddit users may quote other comments simply to emphasise a point or make a joke.

5.5 Performance Discussion

5.5.1 Difficulty in Classifying Pro-Brexit

Throughout the experiments it was observed that our models had significant difficulty with predicting the pro-Brexit stance compared to the other stances. As pro-Brexit constituted less than 6% of the final dataset there may not be enough training instances for our models to get an accurate representation of pro-Brexit stance. If we recall the stance detection results from Table 3.1, we observe that workers have significantly more difficulty annotating for pro-Brexit stance compared to the others. This characteristic of pro-Brexit texts may extend into the performance of our stance detection models suggesting an inherent difficulty in classifying pro-Brexit stance.

5.5.2 BERT Instability

Another performance related issue we observed was high variance in evaluation scores during the training of BERT models. This occurred even when BERT is trained with the same hyperparameters and data and resulted in variations of around 0.05 F1 Score. The literature suggests that BERT fine-tuning is an unstable process [Mosbach

et al., 2020]. Several reasons have been put forth for this issue, namely, catastrophic forgetting, small size of the fine-tuning dataset, optimisation and generalization.

Catastrophic forgetting refers to the tendency for a neural network trained on two different tasks sequentially to forget information learned during the first task when training for the second task. Empirical studies have found this phenomena highly correlated with failure to converge in fine-tuning BERT, however Mosbach et al. [2020] contend that this is rather due to an optimization problem causing catastrophic failure.

Although multiple studies relate the stability of BERT to the size of the dataset, experiments conducted by Mosbach et al. [2020] suggest that this is more so actually due to the number of training iterations further blaming optimization for instability. The authors show that failed fine-tuning runs suffer from vanishing gradients and suggest bias correction or longer training with lower learning rates as solutions. The authors attribute any remaining BERT instability to generalization and suggest that it is even advantageous to train for even more iterations until loss reaches almost zero to over come this issue.

This is consistent with observations made during randomized hyperparameter tuning, which tended to yield low learning rates around 2.5e-5 and long training periods of around 12 epochs compared to the hyperparameters set in the original BERT paper. However this does not explain the remaining instability of the model. It is possible that our method already achieves the highest possible convergence or there may be other unknown reasons for the instability.

Chapter 6

Conclusion

In this thesis, we produce a labelled dataset and build a state of the art stance detection model for Brexit discussions on Reddit to help us better understand the nuances of stance detection on online social media platforms and illuminate the way towards a less polarised society.

We find that a large portion of workers on the MTurk platform do not produce adequate quality annotations and inbuilt qualifications are insufficient for barring low quality workers. This necessitates the use of manual quality vetting measures for workers to achieve high quality labels. We find that it can be particularly difficult for crowdsourced workers to annotate minority stances when dealing with polarised subjects and that this difficulty also extends to machine learning models making predictions on minority classes. Our results suggest in-domain fine-tuning as the optimal strategy for improving BERT performance. However, despite state of the art performance with BERT based models, we find that fine-tuning is often unstable and hyperparameter tuning difficult when resources are limited.

In the following chapter we summarise the contributions made in this thesis and outline directions for future research.

6.1 Summary

We present the following contributions:

- A pipeline for labelling r/Brexit discussions for stance using MTurk.
- An empirical evaluation of different approaches for selecting workers on MTurk to produce the highest quality labelling.
- A dataset of 5895 labelled r/Brexit submissions.
- An analysis of the dynamics around r/Brexit discussions, where we relate the shifts in user activity with political events occurring over the period.
- An evaluation of several novel strategies for improving BERT performance, namely, BERTweet, in-domain pre-training and multi-task fine-tuning with DINO.

- A state of the art stance detection model for Brexit discussions on Reddit achieving a test F1 Score of 0.5547 and external validation F1 Score of 0.6691.

6.2 Future Work

There are several possible avenues for future research which we were not yet able to investigate in this thesis. In the following section we will discuss these options.

6.2.1 Annotation Stance

In our research we found that IAA varies between stance classes. Further investigating the causes and solutions to this issue would be useful for improving the quality of data annotation on crowdsourcing platforms. Additionally such research may give us insights into what makes the expression of pro-Brexit stance unique. One possible option could be to conduct content analysis on the labelled data to determine how language and communication differs between the stances.

6.2.2 Stance Detection

We have yet to investigate the possibility of using Reddit network features to improve stance detection performance. Although Reddit is somewhat limited in this respect, a viable option may be to record the subreddits a user participates in to extract features related to a user's community network. We could also record which other users a user replies to, enabling us to extract features related to a user's interaction network.

6.2.3 Improving BERT

There is a significant amount of work that could be done to improve performance of BERT on our stance detection task. Further experimentation and evaluation is necessary to confirm the effectiveness of our in-domain pre-training approach. This might involve extending the pre-training to cover a larger Brexit discussion corpus by broadening the scope of subreddits to include subreddits such as r/ukpolitics, r/LabourUK or r/tories etc. Additionally, we could also add NSP to our in-domain pre-training to see if it helps improve performance.

Another research avenue we could pursue would be to pre-train a BERT model from scratch using entirely Reddit data, analogous to the approach taken by BERTweet with Twitter. Such a model would be a useful resource for fine-tuning BERT on any downstream NLP tasks which involve more online discussion oriented language.

Bibliography

Turkopticon. https://turkopticon.net/. Accessed: 2021. (cited on page 13)

Ahler, D. J.; Roush, C. E.; and Sood, G., 2019. The micro-task market for lemons: Data quality on amazon's mechanical turk. In *Meeting of the Midwest Political Science Association*. (cited on page 14)

Aldayel, A. and Magdy, W., 2019. Your stance is exposed! analysing possible factors for stance detection on social media. *Proceedings of the ACM on Human-Computer Interaction*, 3, CSCW (2019), 1–20. (cited on page 5)

AlDayel, A. and Magdy, W., 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58, 4 (2021), 102597. (cited on pages 2 and 13)

Amazon Mechanical Turk. Amazon mechanical turk. https://www.mturk.com/. Accessed: 2021. (cited on page 11)

Amazon Web Services, 2021. Amazon mechanical turk best practices. https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanicalTurkRequester/IntroBestPractices.html. Accessed: 2021. (cited on page 14)

Auxier, B. and Anderson, M., 2021. Social media use in 2021. *Pew Research Center*, (2021). (cited on page 7)

Bozdag, E., 2013. Bias in algorithmic filtering and personalization. *Ethics and information technology*, 15, 3 (2013), 209–227. (cited on page 1)

Bruns, A., 2017. Echo chamber? what echo chamber? reviewing the evidence. In *6th Biennial Future of Journalism Conference (FOJ17)*. (cited on page 1)

Callison-Burch, C. and Dredze, M., 2010. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, 1–12. (cited on pages 11 and 16)

Chmielewski, M. and Kucker, S. C., 2020. An mturk crisis? shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11, 4 (2020), 464–473. (cited on page 11)

Conover, M. D.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Menczer, F.; and Flammini, A., 2011. Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*. (cited on page 1)

Dawson, N.; Rizoiu, M.-A.; Johnston, B.; and Williams, M. A., 2019. Adaptively selecting occupations to detect skill shortages from online job ads. In *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, 1637–1643. IEEE, Los Angeles, CA, USA. doi:10.1109/BigData47090.2019.9005967. http://arxiv.org/abs/1911.02302https://ieeexplore.ieee.org/document/9005967/.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, (2018). (cited on page 6)

Eisenstein, J., 2013. What to do about bad language on the internet. In *Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, 359–369. (cited on page 20)

Finkel, E. J.; Bail, C. A.; Cikara, M.; Ditto, P. H.; Iyengar, S.; Klar, S.; Mason, L.; McGrath, M. C.; Nyhan, B.; Rand, D. G.; et al., 2020. Political sectarianism in america. *Science*, 370, 6516 (2020), 533–536. (cited on page 2)

Gallicchio, C.; Martín-Guerrero, J. D.; Micheli, A.; and Soria-Olivas, E., 2017. Randomized machine learning approaches: Recent developments and challenges. In *ESANN*. (cited on page 20)

Ghosh, S.; Singhania, P.; Singh, S.; Rudra, K.; and Ghosh, S., 2019. Stance detection in web and social media: a comparative study. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 75–87. Springer. (cited on page 6)

Heltzel, G. and Laurin, K., 2020. Polarization in america: two possible futures. *Current opinion in behavioral sciences*, 34 (2020), 179–184. (cited on page 2)

Issa, F.; Monticolo, D.; Gabriel, A.; and Mihăiţă, A., 2014. An intelligent system based on natural language processing to support the brain purge in the creativity process. *IAENG International Conference on Artificial Intelligence and Applications (ICAIA'14) Hong Kong*, (Mar. 2014).

Joseph, K.; Friedland, L.; Hobbs, W.; Tsur, O.; and Lazer, D., 2017. Constance: Modeling annotation contexts to improve stance classification. *arXiv preprint arXiv:1708.06309*, (2017). (cited on page 14)

Kong, Q.; Booth, E.; Bailo, F.; Johns, A.; and Rizoiu, M.-A., 2021. Slipping to the extreme: A mixed method to explain how extreme opinions infiltrate online discussions. *arXiv preprint arXiv:2109.00302*, (2021). (cited on page 28)

Kong, Q.; Rizoiu, M.-A.; Wu, S.; and Xie, L., 2018. Will This Video Go Viral: Explaining and Predicting the Popularity of Youtube Videos. In *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*, 175–178. ACM Press, Lyon, France. doi:10.1145/3184558.3186972. https://arxiv.org/abs/1801.04117http://dl.acm.org/citation.cfm?doid=3184558.3186972.

Kong, Q.; Rizoiu, M. A.; and Xie, L., 2020. Describing and Predicting Online Items with Reshare Cascades via Dual Mixture Self-exciting Processes. In *International Conference on Information and Knowledge Management, Proceedings*, 645–654. ACM, New York, NY, USA. doi:10.1145/3340531.3411861. https://arxiv.org/pdf/2001.11132.pdfhttps://dl.acm.org/doi/10.1145/3340531.3411861.

Küçük, D. and Can, F., 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53, 1 (2020), 1–37. (cited on page 5)

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, (2019). (cited on page 20)

Mao, T.; Mihaita, A.; and Cai, C., 2019. Traffic signal control optimisation under severe incident conditions using genetic algorithm. *Proc. of ITS World Congress (ITSWC 2019), Singapore*, (Oct. 2019).

Mardale, A., 2019. *Information Diffusion in Online Communities*. Master's thesis, Universite Jean Monnet, Saint-Étienne, France. (cited on pages 6, 9, and 25)

McCoy, J.; Rahman, T.; and Somer, M., 2018. Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. *American Behavioral Scientist*, 62, 1 (2018), 16–42. (cited on page 2)

Mihaita, A.; LI, H.; and Rizoiu, M., 2020. Traffic congestion anomaly detection and prediction using deep learning. doi:arXiv:2006.13215.

Mihaita, A. S.; Benavides, M.; Camargo, C.; and Cai, C., 2019a. Predicting air quality by integrating a mesoscopic traffic simulation model and air pollutant estimation models. *International Journal of Intelligent Transportation System Research (IJITSR)*, 17, 2 (2019), 125–141. doi:DOI:10.1007/s13177-018-0160-z. https://link.springer.com/article/10.1007/s13177-018-0160-z.

Mihaita, A. S.; Dupont, L.; Cherry, O.; Camargo, M.; and Cai, C., 2018. Air quality monitoring using stationary versus mobile sensing units: a case study from lorraine, france. *Proc. of ITS World Congress (ITSWC 2018), Copenhagen, Denmark*, (Sep. 2018).

Mihaita, A.-S.; Li, H.; He, Z.; and Rizoiu, M.-A., 2019b. Motorway Traffic Flow Prediction using Advanced Deep Learning. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 1683–1690. IEEE, Auckland, New Zealand. doi:10.1109/ITSC.2019.8916852. https://ieeexplore.ieee.org/document/8916852/.

Mihaita, A.-S.; Liu, Z.; Cai, C.; and Rizoiu, M.-A., 2019c. Arterial incident duration prediction using a bi-level framework of extreme gradient-tree boosting. In *Proceedings of the 26th ITS World Congress*, 1–12. Singapore. http://arxiv.org/abs/1905.12254.

Mihăiţă, A.; Camargo, M.; and Lhoste, P., 2014. Evaluating the impact of the traffic reconfiguration of a complex urban intersection. *10th International Conference on Modelling, Optimization and Simulation (MOSIM 2014), Nancy, France, 5-7 November 2014*, (Nov. 2014).

Mihăiţă, A. S.; Tyler, P.; Menon, A.; Wen, T.; Ou, Y.; Cai, C.; and Chen, F., 2017. An investigation of positioning accuracy transmitted by connected heavy vehicles using dsrc. *Transportation Research Board - 96th Annual Meeting, Washington, D.C.*, (Jan. 2017).

Mihăită, S. and Mocanu, S., 2011. An energy model for event-based control of a switched integrator. *IFAC Proceedings Volumes*, 44, 1 (2011), 2413–2418. doi:https://doi.org/10.3182/20110828-6-IT-1002.02082. https://www.sciencedirect.com/science/article/pii/S1474667016439741. 18th IFAC World Congress.

Mishra, S.; Rizoiu, M.-A.; and Xie, L., 2018. Modeling Popularity in Asynchronous Social Media Streams with Recurrent Neural Networks. In *International AAAI Conference on Web and Social Media (ICWSM '18)*, 1–10. Stanford, CA, USA. https://arxiv.org/pdf/1804.02101.pdf.

Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C., 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 31–41. (cited on pages 5, 11, 12, 16, and 17)

Monticolo, D. and Mihăiţă, A., 2014. A multi agent system to manage ideas during collaborative creativity workshops. *International Journal of Future Computer and Communication (IJFCC)*, 3, 1 (Feb. 2014), 66–70. doi:10.7763/IJFCC.2014.V3.269.

Mosbach, M.; Andriushchenko, M.; and Klakow, D., 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*, (2020). (cited on pages 29 and 30)

National Computer Infrastructure, 2021. Hpc systems. https://nci.org.au/our-systems/hpc-systems. Accessed: 2021. (cited on page 21)

Nguyen, D. Q.; Vu, T.; and Nguyen, A. T., 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*, (2020). (cited on pages 20 and 21)

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S., 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32* (Eds. H. Wallach; H. Larochelle; A. Beygelzimer; F. d'Alché-Buc; E. Fox; and R. Garnett), 8024–8035. Curran Associates, Inc. http://papers.neurips.cc/paper/

9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf. (cited on page 20)

/r/datasets mod team, 2019. Pushift reddit api. https://pushshift.io/. Accessed: 2021. (cited on page 9)

Rizoiu, M. A. and Velcin, J., 2011. Topic extraction for ontology learning. In *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances* (Eds. W. Wong; W. Liu; and M. Bennamoun), 38–60. IGI Global. ISBN 9781609606251. doi:10.4018/978-1-60960-625-1.ch003. http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-60960-625-1.ch003.

Rizoiu, M.-A. and Xie, L., 2017. Online Popularity under Promotion: Viral Potential, Forecasting, and the Economics of Time. In *International AAAI Conference on Web and Social Media (ICWSM '17)*, 182–191. Montréal, Québec, Canada. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15553https://arxiv.org/pdf/1703.01012.pdf.

Rizoiu, M. A.; Xie, L.; Caetano, T.; and Cebrian, M., 2016. Evolution of privacy loss in wikipedia. In *WSDM 2016 - Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, 215–224. ACM, ACM Press, New York, New York, USA. doi:10.1145/2835776.2835798. http://dl.acm.org/citation.cfm?doid=2835776.2835798http://arxiv.org/abs/1512.03523http://dx.doi.org/10.1145/2835776.2835798.

Schick, T. and Schütze, H., 2021. Generating datasets with pretrained language models. *arXiv preprint arXiv:2104.07540*, (2021). (cited on pages xiii, 21, and 22)

Shafiei, S.; Mihaita, A.; Nguyen, H.; Bentley, C. D. B.; and Cai, C., 2020. Short-term traffic prediction under non-recurrent incident conditions integrating data-driven models and traffic simulation. In *Transportation Research Board (TRB) 99th Annual Meeting, Washington D.C.* doi:http://hdl.handle.net/10453/138721.

Shafiei, S.; Mihăiţă, A.-S.; Nguyen, H.; and Cai, C., 2022. Integrating data-driven and simulation models to predict traffic state affected by road incidents. *Transportation Letters*, 14, 6 (2022), 629–639. doi:10.1080/19427867.2021.1916284. https://doi.org/10.1080/19427867.2021.1916284.

Snow, R.; O'connor, B.; Jurafsky, D.; and Ng, A. Y., 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, 254–263. (cited on page 12)

Soliman, A.; Hafer, J.; and Lemmerich, F., 2019. A characterization of political communities on reddit. In *Proceedings of the 30th ACM conference on hypertext and Social Media*, 259–263. (cited on page 8)

Sun, C.; Qiu, X.; Xu, Y.; and Huang, X., 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, 194–206. Springer. (cited on page 21)

TurkerView, 2021. Turkerview. https://turkerview.com/. Accessed: 2021. (cited on page 13)

Turner, M. A. and Smaldino, P. E., 2018. Paths to polarization: How extreme views, miscommunication, and random chance drive opinion dynamics. *Complexity*, 2018 (2018). (cited on page 1)

Unwin, J. T.; Routledge, I.; Flaxman, S.; Rizoiu, M. A.; Lai, S.; Cohen, J.; Weiss, D. J.; Mishra, S.; and Bhatt, S., 2021. Using hawkes processes to model imported and local malaria cases in near-elimination settings. *PLoS Computational Biology*, 17, 4 (apr 2021), e1008830. doi:10.1371/JOURNAL.PCBI.1008830. http://medrxiv.org/content/early/2020/07/17/2020.07.17.20156174.abstracthttps://dx.plos.org/10.1371/journal.pcbi.1008830.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I., 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008. (cited on page 6)

Wen, T.; Mihăiţă, A.-S.; Nguyen, H.; Cai, C.; and Chen, F., 2018. Integrated incident decision-support using traffic simulation and data-driven models. *Transportation Research Record*, 2672, 42 (2018), 247–256. doi:10.1177/0361198118782270. https://doi.org/10.1177/0361198118782270.

Wojcieszak, M. E. and Mutz, D. C., 2009. Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement? *Journal of communication*, 59, 1 (2009), 40–56. (cited on page 1)

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., 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Association for Computational Linguistics, Online. https://www.aclweb.org/anthology/2020.emnlp-demos.6. (cited on page 20)

Wu, S.; Rizoiu, M.-A.; and Xie, L., 2019. Estimating Attention Flow in Online Video Networks. *Proceedings of the ACM on Human-Computer Interaction*, 3, CSCW (nov 2019), 1–25. doi:10.1145/3359285. http://dl.acm.org/citation.cfm?doid=3371885.3359285.

Wu, S.; Rizoiu, M. A.; and Xie, L., 2020. Variation across scales: Measurement fidelity under Twitter data sampling. In *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020*, 715–725. https://arxiv.org/abs/2003.09557.

Zhang, R.; Walder, C.; and Rizoiu, M.-A., 2020. Variational Inference for Sparse Gaussian Process Modulated Hawkes Process. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 04 (apr 2020), 6803–6810. doi:10.1609/aaai.v34i04.6160. http://arxiv.org/abs/1905.10496https://aaai.org/ojs/index.php/AAAI/article/view/6160.