

# Beyond Hostility: Detecting Subtle and Overt Forms of Online Conflict with Multi-Objective Learning

**Oliver Warke**

O.WARKE.1@RESEARCH.GLA.AC.UK

*GAIR Lab – School of Computing Science, University of Glasgow  
Glasgow, United Kingdom*

**Joemon M Jose**

JOEMON.JOSE@GLASGOW.AC.UK

*GAIR Lab – School of Computing Science, University of Glasgow  
Glasgow, United Kingdom*

**Jan Breitsohl**

JAN.BREITSOHL@GLASGOW.AC.UK

*Adam Smith Business School, University of Glasgow  
Glasgow, United Kingdom*

**Editor:** Daniela Godoy

## Abstract

Social networks have become a dominant and influential aspect of modern society, with numerous users engaging in observing, creating, and distributing content. The growth of content has led to user conflicts that include bullying, aggression, harassment, and threats. Consequently, recent research has been aimed at identifying and addressing these openly hostile forms of social conflict. However, in current studies, the less overtly hostile yet equally damaging types of conflict, including teasing, criticism, and sarcasm, have been overlooked. We utilise a previously introduced multi-class conflict dataset and extend its application through new modelling approaches for conflict classification, developing a robust multi-objective classification model to capture the full spectrum of conflict; from subtle tensions to open hostility, significantly advancing conflict detection capabilities. This innovative approach leverages class-based reward functions to improve model performance and is implemented by fine-tuning pre-trained transformer language models within a decision transformer framework. We also propose and evaluate multiple knowledge distillation strategies that compress large LLM teachers into efficient student models. These methods result in statistically significant increases in classification performances, leading to further evaluation of performance-cost trade-off. Our experiments on three datasets demonstrate superior recall, precision, F1-score, and accuracy compared to traditional state-of-the-art deep learning classifiers. Furthermore, we analyse class ambiguity and its impact on model performance as well as conducting thematic analysis on model misclassifications.

**Keywords:** Subtle and Overt Hostility, Online Conflict Detection, Social Media Behaviour Analysis, Knowledge Distillation, Transformer-Based NLP

## 1 Introduction

Social networks have become a primary avenue for interpersonal communication, with extensive research and evidence supporting its widespread use within society (Auxier and Anderson, 2021; Ortiz-Ospina and Roser, 2023; Bucher et al., 2018). Social media users communicate opinions, views, and experiences to a broad global audience. Consequently,

increasing levels of negative behaviours and interactions have been observed on social media platforms (Akram and Kumar, 2017; Siddiqui et al., 2016; Berryman et al., 2018). Several factors contribute to the escalation of these behaviours, including anonymity, the absence of physical and social cues, and limited accountability (Gonzales, 2014). These detrimental conflict behaviours take many forms and are constantly evolving to present different threats. All of these factors lead to a toxic online environment, with severe consequences to the well-being of individuals and communities.

Existing research methodologies aim to detect overtly hostile and socially damaging behaviours, such as hate speech, aggression, and cyberbullying, primarily due to their evident nature (Fortuna and Nunes, 2018; Alkomah and Ma, 2022; Poletto et al., 2021). However, subtle negative behaviours, for example, sarcasm and teasing, have been shown to exert a significant detrimental influence on user well-being (Boroon et al., 2021; Kowalski, 2000; Wang et al., 2022; Ledley et al., 2006), since their perceived harm varies subjectively. We argue that lighter forms of conflict—such as criticism, sarcasm, and teasing—exist along a broader spectrum of negative behaviours. This spectrum also includes more severe actions like harassment and threats, which have received comparatively greater research attention. For example, all user behaviours depicted in Poletto et al. (2021), are located at the end of the spectrum. Nonetheless, little research focuses on detecting broader manifestations of conflicts, and hence it becomes imperative to investigate a diverse spectrum of adverse user interaction behaviours on social media, collectively referred to as "conflicts". We argue that lighter forms of conflict are not merely harmless interactions; they can independently exert significant negative effects on users and may also contribute to the escalation of exchanges into more severe and overtly hostile behaviours. Prior research has shown that negative interactions in online environments can intensify over time and develop into more hostile forms of communication (Cheng et al., 2015; Kowalski et al., 2014). This gap highlights the need for more nuanced modelling frameworks capable of capturing the full spectrum of conflict behaviours.

Detecting and addressing both the lighter and more extreme forms of negative behaviour along our conflict spectrum would allow for the behaviours to be addressed before escalating. Limiting detection to only the more overt forms of negative behaviour results in reactionary steps, rather than proactive prevention or monitoring. As discussed above, terminology within the domain tends to centre around more extreme negative behaviours. Bullying is one such commonly referenced term. However, whilst we do not diminish the impact of bullying, its core component of repetition is not necessary for harm to be inflicted on users. We argue that even a single instance of negative behaviour can significantly harm users, also supported by Wolak et al. (2007). These limitations highlight the need for nuanced conflict modelling frameworks and classifiers detecting a broader spectrum of individual conflict occurrences, which are scalable, efficient, and effective. This challenge is further compounded by recent shifts in platform-level moderation, where major social media sites are increasingly delegating content oversight to users. This raises urgent questions about how conflict detection can be scaled and automated.

It is also important to recognise how negative behaviours, negative user interactions, and negative content on social media are all becoming more commonplace and accepted. Whilst it is important to strike a balance between free speech and moderation of content, the acceptance that users will be subjected to large volumes of negative content is disturbing.

Popular sites such as Facebook and X (Twitter) are abandoning their more strict content policies in favour of more lax content controls. In January 2025, Mark Zuckerberg, the CEO of Facebook, announced that Facebook would be removing fact-checkers and moderators. Zuckerberg accepted that there would be more misinformation and hate on the platform but that it would be a worthwhile trade-off (Thompson and Conger, 2025; Duffy, 2025; Liv McMahon, 2025). In a video released on Meta, Zuckerberg stated that they were going to rely on users reporting content violations before they took action (Meta and Kaplan, Joel, 2025; Zuckerberg, 2025). This means that although many users can be exposed to harmful content on the platform, Meta will not take action unless users report it, despite having the ability to detect such content with their current system. X (Twitter) has also removed much of its content moderation team, which has also caused concern about the increase in harmful content on the platform (Alba and Wagner, 2023; Brewster, 2024; Sankaran, 2022). This in turn can only have damaging results for users on these platforms as they face increased levels of negativity, toxicity, and harmful content. Both Facebook and X have introduced "community notes" as a form of fact checking and content moderation, these notes place the burden on the users to moderate and fact-check the content on the platform, with little oversight from the platform and no training. This extra responsibility can only serve to expose users to even more negative content as they are forced to engage with the harmful content in an attempt to preserve a positive environment. This increasing relaxation of content policies not only reinforces the need to develop measures to detect negative user conflicts but also the need to raise awareness of the potential epidemic of negative social media spaces. Although it is always important to promote free speech and fight against censorship, it does not mean that users' mental and physical health should be ignored and moral principles abandoned.

With the increasing recognition of conflicts as a significant research area and a growing social issue (Fortuna and Nunes, 2018), there has been a rise in the availability of datasets and models focused on the phenomenon (Matamoros-Fernández and Farkas, 2021; Aboujaoude et al., 2015). Existing research has centred on binary hate datasets and published papers, with fewer multi-label and multi-class datasets available (Davidson et al., 2017; Elsafoury, 2020; Mollas et al., 2022). While these resources have contributed valuable insights, it is essential to note that current research datasets have limitations and do not encompass a wide range of social media conflicts. Hence, our work utilises a multi-class conflict dataset, developed using hate netnographies (Kozinets, 2002) and introduced in a previous work (Warke et al., 2024). This dataset serves as a benchmark for the evaluation of models specifically within a nuanced human behaviour task. When used in combination with other multi-class social media hate and negative behaviour datasets, it forms a thorough experimental setup for rigorously and systematically testing classification models across multiple domain adjacent tasks.

In multi-class classification, a substantial challenge arises due to the inherent interaction between distinct classes, leading to heightened complexities in distinguishing and isolating specific class instances. Particularly within multi-class scenarios, the occurrence of cross-talk between classes introduces a significant hurdle, consequently impacting the effectiveness of classification (Lango and Stefanowski, 2022). This challenge is worsened in social media scenarios due to the ambiguity, noisy and error-prone constitution of social media data (Bouazizi and Ohtsuki, 2019). In addition, the conventional strategy of

constructing test collections from social media, especially exploiting distant supervision, contributes to this situation’s intricacies. In hate classification research, despite other researchers identifying the complexity and ambiguity of defining these behaviours, there has been no discussion on the profound impact of these issues on classification performance. To address these challenges, we evaluate three modeling strategies; Decision Transformer, hierarchical classification, and knowledge distillation. Each of these methods is designed to mitigate different aspects of multi-class ambiguity.

As a use case we investigate detecting conflicts effectively and experiment with social media data in prominent consumer brand communities. Brand communities are pages, groups, and timelines controlled by brands which provide the opportunity for consumers to interact with the brand and each other (Brogi, 2014). These communities embrace substantial global user cohorts, thereby rendering them of notable significance for consumer conflict research. Prior work within the marketing domain has underscored the challenges brands encounter in relation to conflicts. These challenges include but are not limited to; serious damage to brand reputation, harm to consumers well-being, withdrawal of users from the brand community, and consumer purchase intentions (Breitsohl et al., 2018; Wang et al., 2012).

Although the dataset is derived from brand community interactions, these environments represent an important and widely studied form of social media interaction. Brand communities host large and diverse groups of users who engage in discussions with both brands and other users, generating a wide range of communicative behaviours including disagreement, criticism, and conflict. As such, they provide rich and naturally occurring data for analysing online interactions. The behavioural categories captured in this dataset, such as teasing, criticism, sarcasm, trolling, harassment, and threats, are not unique to brand communities and have been documented across many social media platforms. Nevertheless, we acknowledge that the dataset originates from a specific online context. While the behaviours themselves are broadly observed in online discourse, further evaluation across additional platforms would strengthen the generalisability of the proposed modelling approaches.

As discussed, current approaches fail to capture the spectrum of conflicts, focusing research on methods that detect only extreme forms of hate. Therefore, in this work, we investigate three methods for increasing the baseline performance of a classification model: a Reward-based Model called ConflictDT, Hierarchical Classification Framework, and Knowledge Distillation. ConflictDT, draws inspiration from recent advances in sequential decision modeling, where class-level signals are used to guide model behaviour. It was first introduced in this prior work (Warke et al., 2024). The Hierarchical framework leverages relationship structure within conflict types, enabling coarse-to-fine classification across grouped behaviours. Finally, Knowledge distillation compresses large language models into efficient student models, allowing us to explore trade-offs between model size, supervision depth, and classification accuracy.

There are a few significant differences between ConflictDT and Knowledge Distillation approaches which makes for an interesting comparison; whilst both seek to improve the performance of a baseline model using additional signals, ConflictDT does so without the need for an expensive powerful LLM acting as a teacher. The evaluation of the two approaches centres on the relationship between performance increase and cost. Whilst other works do

touch upon this relationship comparing the baseline model with the enhanced model, few conduct a detailed performance-cost analysis of different enhanced model approaches.

We make the following contributions in this paper:

- We present a comparative evaluation with statistical analysis of three modeling strategies—ConflictDT, hierarchical classification, and knowledge distillation for improving baseline performance in nuanced conflict classification and other tasks.
- We adapt and evaluate a hierarchical classification model to exploit the relationship structure within conflict class groupings.
- We propose and compare three distinct knowledge distillation strategies for compressing LLMs into efficient student models; Soft Target Distillation, Feature Distillation using Probabilities, and Feature Distillation using Hidden State Embeddings.
- We conduct a performance-cost analysis of ConflictDT and KD approaches, quantifying trade-offs between accuracy and computational efficiency.
- We conduct a comprehensive empirical analysis of distillation hyperparameters across various benchmarks to evaluate performance, scalability, and efficiency.
- We analyse classifier performance beyond traditional metrics, conducting a popular social science technique, thematic analysis, to investigate misclassification trends.

Building on these contributions, we define the following research questions to structure our investigation:

- R.Q.1 How effective are the different classification models for improving multi-class classification? How do the different forms of knowledge distillation compare in performance?
- R.Q.2 Can the reward function within ConflictDT be used to guide model behaviour and improve classification performance?
- R.Q.3 How effective is the Hierarchical Model in capturing latent structure within conflict class groupings, and how does it compare to ConflictDT when distinguishing between lesser and more extreme conflict behaviours?
- R.Q.4 What patterns emerge in model misclassifications, and how do these align with social science theories of online communication?
- R.Q.5 How do distillation hyperparameters affect the performance, scalability, and efficiency of each knowledge distillation strategy?
- R.Q.6 How do model training costs compare with classification performance across baseline, ConflictDT, and KD-enhanced models?

## 2 Related Work

### 2.1 Hate and Conflict Classification

There are many works investigating hate detection that have developed various machine-learning models and explored different classification algorithms and feature representations. Advancements in deep learning have increased the effectiveness significantly (Minaee et al., 2021). Deep learning algorithms, including BERT (Devlin et al., 2019), LSTM (Hochreiter and Schmidhuber, 1997), and CNN (O’Shea and Nash, 2015), have demonstrated strong performance across different text classification tasks. Khan et al. (2020) proposed a CNN-based framework called HateClassify for labelling social media content, achieving competitive multiclass accuracy. Researchers have explored techniques to enhance the performance of these algorithms further (d’Sa et al., 2020; Naseem et al., 2019). Ali et al. (2023) developed a LSTM-GRU model, combining deep learning and graph analytics, which exhibits state-of-the-art performance over a six class hate speech dataset.

The Transformer architecture (Vaswani et al., 2017), commonly utilised in language modelling and recommendation tasks (Li et al., 2020; Sun et al., 2019a), has revolutionised text classification tasks. This is attributed to its capacity to capture relational information using self-attention mechanisms. Salminen et al. (2020) focused on detecting hate across multiple platforms and found models with BERT features superior. d’Sa et al. (2020) conducted hate speech classification using word embeddings and Deep Neural Networks (DNN), with the fine-tuning approach using BERT achieving significant improvements. Caselli et al. (2021) introduced HateBERT, a retrained BERT model that outperformed the base BERT model in detecting abusive language.

Another closely related line of research focuses on toxicity scoring frameworks used in large-scale content moderation systems. Tools such as Google’s Perspective API estimate the likelihood that a comment will be perceived as toxic, abusive, or otherwise harmful, producing continuous toxicity scores that can be used to prioritise moderation decisions (Wulczyn et al., 2017b; Jigsaw). These systems are widely used in industry settings due to their scalability and simplicity, and toxicity scores are frequently used in research to analyse harmful language and evaluate moderation systems (Gehman et al., 2020; Vidgen et al., 2021b). However, toxicity scoring typically focuses on estimating overall harmfulness rather than distinguishing between specific behavioural categories. As a result, such approaches may struggle to capture the nuanced distinctions between different forms of conflict behaviours, such as teasing, sarcasm, criticism, or trolling. In contrast, this work models conflict as a structured multi-class classification problem, enabling finer-grained differentiation across a spectrum of antagonistic behaviours.

As evidenced by the focus of these influential works, existing detection methods tend to focus on overtly hostile behaviours, such as threats and harassment, while overlooking subtler forms like teasing, sarcasm, and criticism. This is primarily a function of the training collections used, which focus on high-severity language like hate speech and explicit abuse (Salminen et al., 2020; d’Sa et al., 2020; Caselli et al., 2021). This gap motivates the need for multi-class conflict classification frameworks that capture a broader behavioural spectrum.

## 2.2 Reward-Based Modelling in NLP

Recent work has explored the use of reward signals to guide model behaviour in NLP tasks. Decision Transformers (Chen et al., 2021) have shown promise in vision and language modelling domains (Wang et al., 2024, 2025; Lee et al., 2022; Janner et al., 2021), but their application to text classification remains under explored. Gontier et al. (2023) adapted Decision Transformers for interactive text environments, demonstrating that reward signals can be integrated with natural language inputs to influence model decisions. Our work builds on this foundation by applying class-based reward signals to guide multi-class conflict classification.

## 2.3 Hierarchical Classification Models

Hierarchical classification has been widely used in emotion and sentiment detection tasks. Hadjiharalambous et al. (2022) introduced a hierarchical approach within an emotion detection task seeks to first classify datapoints into higher level groups before classifying them into their final classes. The primary emotion task features emotions which fall into two higher level classes, positive and negative. The negative group features three emotions, whilst the positive only one. The paper also introduces secondary emotion detection tasks with multiple emotions in the positive group but does not achieve statistically significant results in these tasks. The novelty of applying the model within this task is therefore threefold. Firstly, the groupings within the conflict task, lesser and more extreme conflict behaviours, are less defined than positive and negative emotions. Therefore, it is interesting to see how the hierarchical model performs with more complex relationships in an adjacent domain. Secondly, although the hierarchical model performed better in the tasks involving multiple sub-groups within each higher level group, the results did not show statistically significant increases in performance. By using the hierarchical model within this paper, we expand upon and enhance the methodology, evaluation, and analysis in the previous hierarchical work.

## 2.4 Knowledge Distillation

Knowledge Distillation (KD), first formally introduced by Hinton et al. (2015), investigated transferring the "dark knowledge" from a large, high-performance "teacher" model to a smaller, more efficient "student" model. This significant research demonstrated that by training the student to mimic the softened output probabilities (often referred to as "soft targets") of the teacher, the student could achieve significantly improved performance compared to the baseline performance of the student model. This initial concept provided a framework for distilling complex decision boundaries, class relationships, and class characteristics that are not easily understood by the student model.

In its early stages, KD involved matching the classification outputs of the student and teacher models. FitNets (Romero et al., 2015) introduced the idea of using intermediate feature representations, termed "hints," to guide student models. This approach allowed the student to learn not just the teacher's final outputs, but also the internal feature representations of the teacher which contained rich detail. Patient Knowledge Distillation (Sun et al., 2019b) refined feature-based transfer by matching hidden representations across mul-

tiple layers of the teacher and student. This multi-layer alignment strategy often leads to improved convergence, more robust feature transfer, and better generalisation for the student model, as it allows the student to learn similar hierarchical representations to the teacher. Zagoruyko and Komodakis (2016) introduced attention mechanism distillation, where the student learns to mimic the teacher’s attention patterns. These diverse approaches have expanded the KD domain and led to its application in many different fields.

The application of KD has been widely explored within transformer models. DistilBERT (Sanh et al., 2019) was a significant example of KD which drastically reduced the size of BERT while maintaining much of its performance. DistilBERT simplified the training pipeline to use output logits and intermediate embeddings with a specific loss function. TinyBERT (Jiao et al., 2020) introduced a multi-faceted approach for KD in transformers, utilising not only output logit distillation but also fine-grained transformer layer and hidden state distillation, demonstrating robust performance and the power of internal LLM representations. These works demonstrate that KD is a powerful and effective tool for improving the performance of smaller transformer models for deployment in real-world scenarios.

With the introduction of generative LLMs and their future iterations like GPT-3 (Brown et al., 2020), FLAN-T5 (Longpre et al., 2023), and LLaMA (Touvron et al., 2023), issues surrounding the scale and resource requirements of these models have become more prominent. Direct deployment is often too expensive or requires too much computational power, leading to many researchers exploring the different approaches to compressing LLMs via distillation and parameter-efficient tuning. Recent surveys, such as Yang et al. (2024), provide a taxonomy of LLM distillation approaches, categorizing them into output, feature, and hybrid-based methods. However, many of these models are evaluated on general tasks, with limited works examining nuanced tasks within specialised domains.

Feature-based knowledge distillation is common among models which are employed in semantically rich domains, where data consists of a multitude of complex factors such as nuance, context, and class relationships—e.g., conflict behaviour text classification. In the context of transformers, matching hidden state embeddings has been shown to be a strong KD approach (Wang et al., 2020b). More recently, representation-aware distillation has been applied to multitask and multilingual models (Turc et al., 2019).

Our work builds on these foundations by comparing multiple KD strategies in a multi-class conflict setting. Specifically, we evaluate soft loss distillation, feature distillation via probabilities, and feature distillation via hidden state embeddings. These methods are contrasted against ConflictDT and hierarchical classification to assess their effectiveness in capturing subtle behavioural distinctions. Despite recent progress, the best approach for distilling teacher model knowledge in nuanced classification remains an open question—one we address through systematic evaluation of loss signals and distillation weights.

### 3 Methodology

**Problem Statement:** We first define the problem as a text classification task with a dataset  $D = (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  where each datapoint contains a text  $x$  and a corresponding true label  $y$ . The discrete action  $y$  taken by the model is the classification

of the data point into the set of classes. Each true label  $y$  belongs to a finite set of classes  $C = 1, \dots, N$  where  $N$  is the number of classes.<sup>1</sup>

### 3.1 ConflictDT classifier:

Our first approach, ConflictDT, reconceptualises multi-class conflict classification as a decision-transformer-based learning problem. Instead of treating classification as a static mapping between input and label, we model it as a sequence of decisions guided by reward signals. Building on the Decision Transformer framework (Chen et al., 2021), we design a distance-aware reward function that quantifies inter-class separation for each input text, thereby mitigating cross-talk between overlapping class boundaries and promoting more distinct and interpretable decision trajectories.

While the Decision Transformer has previously been applied in text-based reinforcement learning scenarios (Gontier et al., 2023), its use for supervised text classification remains largely unexplored. In this work, we extend the framework beyond traditional reinforcement learning by integrating class-based reward computation directly into the classification pipeline. This innovation allows us to exploit the DT’s reward-conditioning capability without relying on an explicit RL environment, effectively transforming a supervised classification task into a structured decision-making problem. The resulting architecture combines raw textual features with class-distance rewards to achieve enhanced discrimination and reasoning coherence across classes.

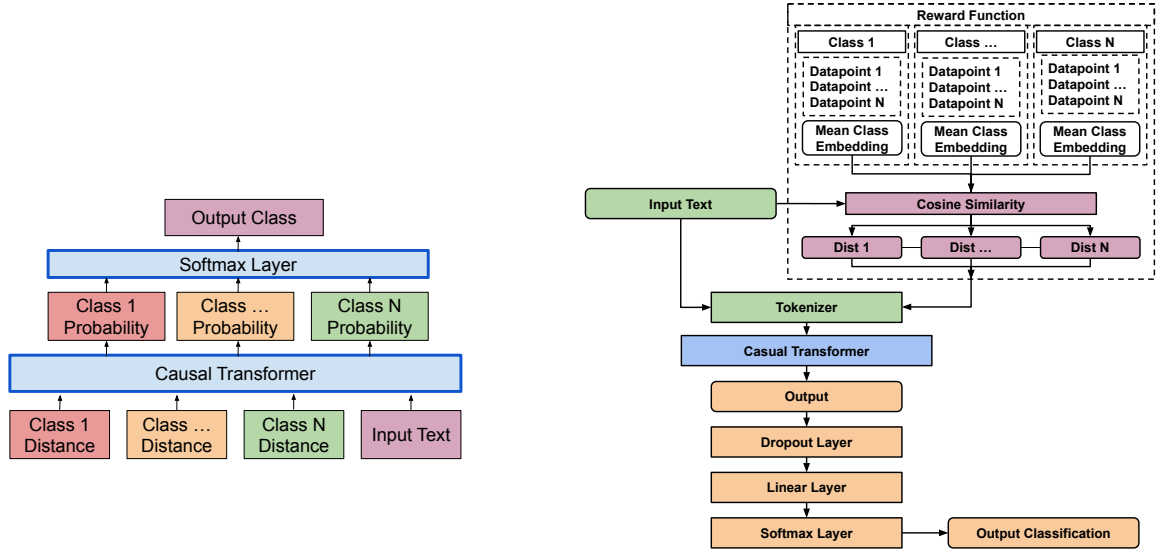


Figure 1: Model Diagrams of ConflictDT.

1. Code and dataset are available at <https://github.com/OliverWarke/ConflictDT>

### 3.2 Class Distance Reward Design

The focal point of this reward function is to optimise class distances with the overarching goal of enhancing classification performance. The core idea of our methodology is to improve upon classification performance when solely using raw datapoint text. We do this by providing the classifier with a quantitative measure of inter-class separation by creating a reward signal that quantifies the separation between the input text and each class. This aims to mitigate challenges due to inherent class ambiguities, inconsistency in short text lengths, and the contextual nature of nuanced human behavioural language, e.g. distinguishing between light teasing and more extreme trolling.

A visual representation of our proposed architectural framework is presented in Figure 3.1. Given an input text  $X$ , the input is first passed through a Sentence Transformer model, separate from the final classifier, to produce a dense text embedding  $e_x$ . This initial encoding is used exclusively to calculate the class distance, providing a robust, non-contextualised feature set.

As a basis for the reward function, we calculate the mean text embedding  $\bar{y}_k$  for each class  $k \in C$  from the training set, where  $C$  is the set of classes within the training set.

$$\bar{y}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} y_{k,i} \quad (1)$$

The reward  $r_{\bar{y}_k e_x}$  for each class  $k \in C$  is computed using cosine similarity between the mean class embedding  $\bar{y}_k$  and the input text embedding  $e_x$ :

$$r_{\bar{y}_k e_x} = \text{CosSim}(\bar{y}_k, e_x) \quad (2)$$

The resulting set of similarities  $\{r_{\bar{y}_k e_x}\}_{k=1}^N$ , i.e. class distances, forms the raw reward vector. We explored two distinct computational methods for employing the cosine similarity metric to quantify inter-class distances. The first method entailed calculating the mean vector for each of the class clusters (1) and gauging the cosine similarities between these mean vectors and text embedding. Conversely, the second method involved computing the cosine similarity between each possible pair of text embeddings of given text and elements of each class, then calculating an average cosine similarity. Upon evaluation, the first approach was deemed more efficient, thus becoming the chosen method implemented within our model.

The similarity/class distance values are then normalised and subsequently scaled to an integer range of 1 to 100. We adopted this scaling decision following research by Wallace et al. (2019), who investigate various NLP model’s understanding of numbers. They found that transformer models do capture numerical features but some, notably BERT, struggle with decimal numbers. By scaling the distances we present them as integers making it more likely that BERT will understand them.

The scaled, normalised set of reward distances is converted into a string token, e.g. [Class Name: Scaled Reward Value,...] and prepended to the raw text input  $x$  to form the final input sequence  $S_{in}$ .

$$S_{in} = [RewardDistanceString|x] \quad (3)$$

Following these steps, the model follows the standard architecture of a deep learning transformer model, shown in Equation 4, with the exception of the input  $x$  which would normally be a textual embedding and is instead a combined text and reward feature embedding. The causal transformer structure of the DT model is simulated using BERT’s self-attention mechanism to learn the relationship between the reward and the subsequent text.

$$z = \sum_{i=1}^n w_i * x_i + b, \quad a = \psi(z) \quad (4)$$

where  $w$ , weight, and  $b$ , bias, are trainable parameters,  $\psi$  is the Softmax activation function, action,  $a$ , is selecting one of the  $N$  dataset classes.

As depicted in Figure 1, the text and reward features are combined before going through the tokeniser and decision transformer layer, which further goes through a number of transformations. The output of the transformer model is first passed through a dropout layer to help prevent overfitting. The dropout layer randomly sets a number of the input tensor elements to zero depending on a set probability. Next, a linear layer, is used to produce the output logits, a vector of raw predictions for each class within the task.

Finally, a softmax layer is then used to produce the final probability distribution  $\mathbf{a}$  over the  $N$  classes from the dataset using the final transformer output  $\mathbf{z}$ . This probability distribution is subsequently used to obtain the predicted class for the datapoint.

$$a = \psi(z) = \frac{e^{z_i}}{\sum_{j=1}^N e^{z_j}} \quad (5)$$

Loss for the model is calculated using cross-entropy loss  $L_{CE}$  as follows:

$$L_{CE} = - \sum_{i=1}^N t_i \log(p_i) \quad (6)$$

for  $N$  classes, where  $t_i$  is the true label and  $p_i$  is the softmax probability for the  $i$ th class.

Although the Decision Transformer paper (Chen et al., 2021), which formed the foundation for this work, utilised the GPT-2 model (Radford et al., 2019), we ultimately decided to use the BERT model (Devlin et al., 2019) as the underlying transformer model within our research. This was due to BERT’s superior performance in initial experiments, a decision reinforced by the findings of the experiments within this paper (see Tables 3, 4, 5, and 6).

In addition to this architectural change to the original work, we introduce a second alternative approach: the novel design of reward functions. Rewards can be modelled by various techniques tailored to suit distinct problem tasks, and we have experimented with different schemes, emphasizing the diversity of our approach. We explored numerous variations of reward functions based on distances during our experiments (section 4).

While the original DT paper (Chen et al., 2021) utilised transformer models within sequence modelling problems, our work presents a novel contribution by extending this framework to text classification. We further extended the framework to text classification by mapping states to sentences within each datapoint. By adapting the DT model to text classification, we enable the model to capture sequential and contextual information within the text datapoint. Subtle human conflict behaviours often depend on emphasis being placed on words or sentences occurring within a sequence. This adaptation and theoretical consideration represents a targeted extension of the original DT framework. To capture the information cues we explore two methods of dividing the text datapoints into sequences. Firstly, we divided the datapoints into sentences, with each state consisting of the previous states plus the next sentence. The first state would contain the first sentence, the second would contain the first two sentences; etc. The model would then provide a classification for each state, with the final state being the entire text datapoint. The second method was dividing the text into 3 sections with equal numbers of words. Then using the same method of combining the states at each timestep. The sequential classification models operate by inputting the current state’s text and the rewards (if being used) to a BERT model. Then if timestep is  $>1$ . The CLS output of the model is combined with the CLS output at the previous timestep. This combined CLS embedding is passed through linear layers to produce a classification decision. This sequence modelling approach aimed to exploit the sequential nature of text, exposing the model to increased contextual information at each timestep, while incorporating the previous timesteps output layers.

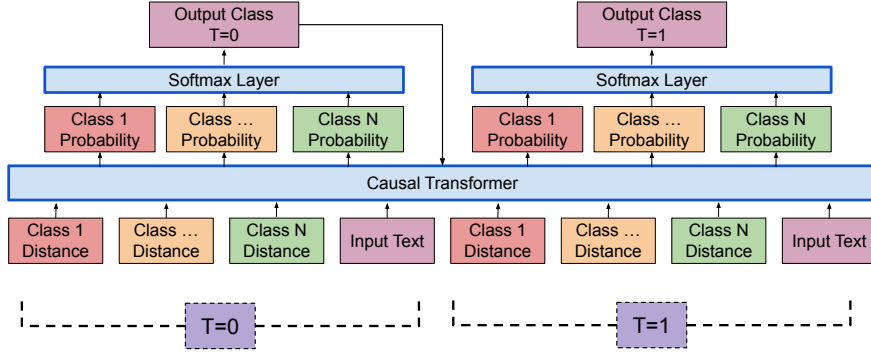


Figure 2: Architecture of Sequential ConflictDT.

### 3.3 Hierarchical Model

The conflict behaviours within our classification task naturally form a hierarchy. The behaviours exist within a range of light (teasing, criticism, sarcasm) to more extreme behaviours (trolling, harassment, threats). This division into parent groups and sub-classes suggests that a hierarchical classification strategy may be an effective means of exploiting these class relational groupings to improve classification performance.

We implement the Hierarchical model as described in Hadjiharalambous et al. (2022), with significant modifications to the methodology. We exploit the contrasting characteristics between the lesser and more extreme conflict behaviours. The Hierarchical version of

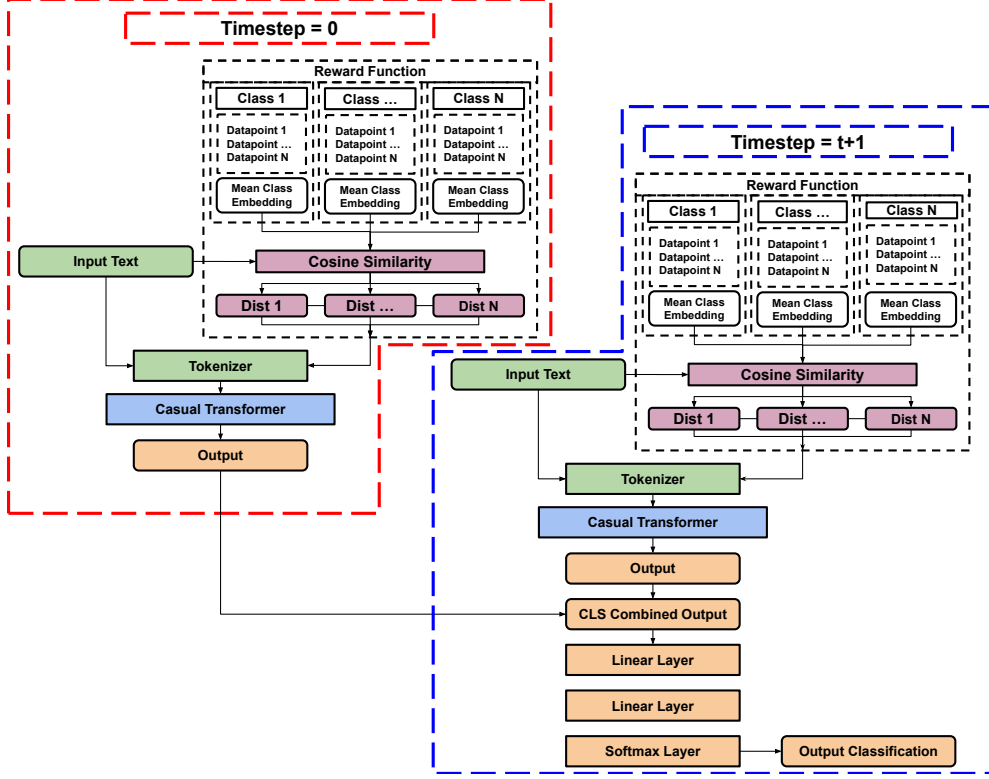


Figure 3: Sequential state construction for ConflictDT.

our classification task is displayed in Figure 4. In their work’s main experiment Hadjiharalambous et al. use an emotion dataset featuring only one positive emotion, and therefore use only two separate loss functions; binary cross-entropy for higher level binary sentiment classification and negative log-likelihood for the multiclass negative emotion classification. This is then represented as:

$$L_T = L_1(\text{Output}_{Bi}) + L_2(\text{Output}_{Nmul}) \quad (7)$$

In our work we have multiple subclasses under each higher level group. We therefore use three separate loss functions; binary cross-entropy for higher level conflict severity classification, and two negative log-likelihood losses for the multiclass lesser and more extreme conflict subclass classification. These are then combined to produce the total model loss, consisting of the three sub-section losses.

$$L_T = L_1(\text{Output}_{Bi}) + L_2(\text{Output}_{Lesser}) + L_3(\text{Output}_{Extreme}) \quad (8)$$

In their methodology Hadjiharalambous et al. include a model architecture featuring a BiLSTM model, however, we opt to use a BERT model within the Hierarchical architecture for two reasons. Firstly, in their paper the BERT based Hierarchical model achieves the

best performance, and secondly we use a BERT model as a baseline model within the rest of the work so it allows for more robust performance comparisons.

This adaption from Hadjiharalambous et al. represents a more nuanced complex application of the hierarchical concept. Within emotions there are clearly defined positive and negative categories whilst conflict behaviours have more subtle differences. Additionally, we have modified the loss function to incorporate multiple sub-classes within the parent groups. Our hierarchical model therefore addresses a domain where the boundaries between parent groups are less defined and each sub-classification task is more complex. This tests the hierarchical model in a novel setting that has not previously been investigated.

Finally, we exploit the relationships between the two behaviour groups within the conflict task as a reward within the ConflictDT model, by utilising the performance of the hierarchical model we can compare and contrast two variations of classification model which seek to exploit the same structural relationships within the task in different ways.

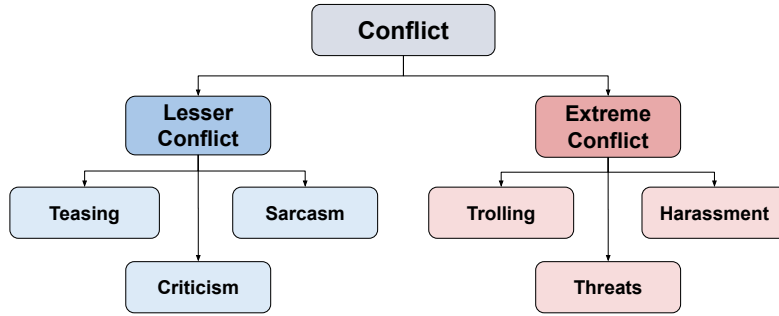


Figure 4: Hierarchy of the Conflict Classes

### 3.4 Distillation Methodologies

To explore the range of distillation signals available from teacher models, we evaluate three representative strategies that vary in granularity and depth of supervision:

- **Soft Target Distillation:** Leveraging softened output distributions from the teacher to guide the student. This is a widely used distillation strategy introduced by Hinton et al. Hinton et al. (2015).
- **Feature Distillation using Probabilities:** Introducing a regression head to align the student’s probability vectors with those of the teacher. This has been shown in prior works to be competitive with soft target distillation Kim et al. (2021).
- **Feature Distillation using Embeddings:** Aligning the student’s internal hidden states with those of the teacher model Romero et al. (2015); Sun et al. (2019b).

Each training instance contributes two learning signals to the student model:

- **Hard label:** The ground truth class provided by human annotators.
- **Soft supervision:** Probability distributions or hidden-state representations extracted from the teacher model through prompted inference.

Using both signals follows standard knowledge distillation practice where the hard labels ensure alignment to the ground truth annotations whilst the soft supervision provides additional information from the teacher which is otherwise not available to the student.

The total loss for all distillation methods is computed as:

$$\mathcal{L}_{\text{total}} = (1 - \alpha) \cdot \mathcal{L}_{\text{CE}} + \alpha \cdot \mathcal{L}_{\text{Distill}} \quad (9)$$

where  $\mathcal{L}_{\text{CE}}$  is the standard cross-entropy loss with hard labels, and  $\alpha$  controls the influence of distillation-specific loss  $\mathcal{L}_{\text{Distill}}$ .

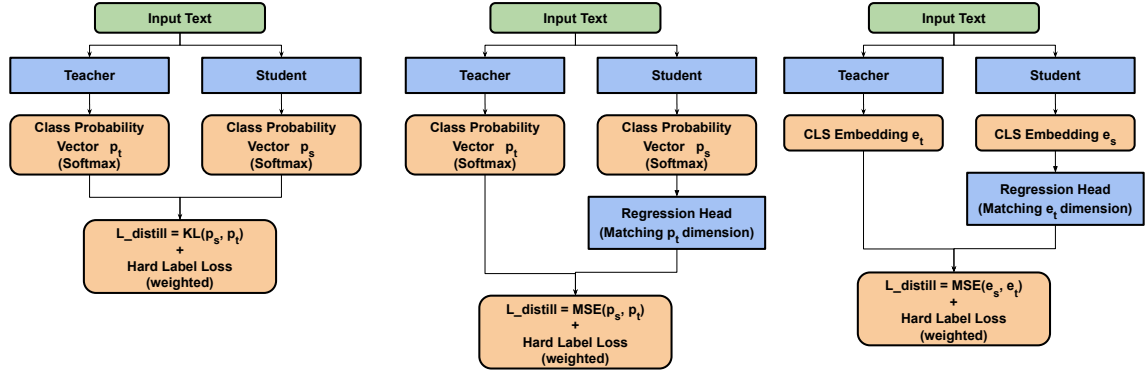


Figure 5: Figures showing Soft Target distillation, Feature Distillation, and Hidden State Embedding Distillation

Although these three distillation strategies are well-established in the knowledge distillation literature and are not novel in themselves, their comparative effectiveness has received limited evaluation in the context of fine-grained conflict classification. Our contribution lies in applying each method to this task, analysing how their differing supervision signals interact with the conflict taxonomy, and conducting an empirical study of their performance across a range of distillation weights. In our distillation experiments, the teacher model is **Meta-Llama-3-8B-Instruct**. Rather than fine-tuning the teacher as a task-specific classifier, we use it in prompted inference mode to generate supervision signals for the student model. All three distillation variants use the same teacher model to ensure comparability between supervision signals.

#### 3.4.1 SOFT TARGET DISTILLATION

In this KD approach (Hinton et al., 2015), the teacher provides a softened probability distribution  $p_t$  over the class set. The student generates its own separate classification distribution  $p_s$ .

$$\mathcal{L}_{\text{Distill}} = \text{KL}(p_t \parallel p_s) \quad (10)$$

This loss captures the fine grained similarities between classes, which are lost in one-hot ground truth labels. See Figure 3.4 for a schematic overview.

### 3.4.2 FEATURE DISTILLATION (PROBABILITIES)

Here, the student is tasked with directly regressing the teacher’s class probability vector  $\mathbf{p}_t$ :

$$\mathcal{L}_{\text{Distill}} = \text{MSE}(\mathbf{p}_s, \mathbf{p}_t) \quad (11)$$

Unlike KL-divergence, MSE treats all errors equally, encouraging the student to match every probability with high fidelity (Kim et al., 2021). We implement this using a separate regression head parallel to the classification head. The architecture is shown in Figure 3.4.

For this conflict classification task, we chose MSE instead of KL-divergence as the conflict behaviours are semantically similar. KL-divergence emphasises relative differences between probabilities, which can amplify noise when teacher models are uncertain. MSE, however, encourages the student to match the teacher’s output for all classes rather than just the relative ranking in KL-divergence. This is suitable to this task as the subtle distinctions between conflict behaviours mean that it is beneficial to the student model to be aware of all class probability not just the dominant class in the rankings.

### 3.4.3 FEATURE DISTILLATION (HIDDEN STATE RAW EMBEDDINGS)

This approach transfers deeper knowledge by aligning internal representations. Specifically, we extract the final layer embedding vector  $\mathbf{e}_t$  from the teacher and train the student to produce a matching embedding  $\mathbf{e}_s$ :

$$\mathcal{L}_{\text{Distill}} = \text{MSE}(\mathbf{e}_s, \mathbf{e}_t) \quad (12)$$

The student uses a regression head that maps its CLS token representation to the same dimensionality as the teacher’s embedding.

## 3.5 Multi-class Conflict Dataset

Building on prior critiques of binary hate speech datasets (Davidson et al., 2017; Khan et al., 2020; Founta et al., 2018), we address the lack of coverage across the full spectrum of conflict behaviours by utilising a previously introduced multi-class conflict dataset derived from Facebook brand community interactions. While existing datasets predominantly focus on overtly hostile behaviours, this dataset captures a broader spectrum of conflict, including subtle forms such as teasing, sarcasm, and criticism.

The dataset originates from prior research investigating consumer conflicts within online brand communities (Breitsohl et al., 2018; Warke et al., 2024). It was constructed using a hate netnography methodology (Kozinets, 2015), which involves the systematic observation and analysis of naturally occurring online interactions. In the original study, social media interactions were collected across multiple brand communities over a sixteen-month period, enabling the identification of recurring patterns of consumer conflict behaviour. The resulting dataset was manually annotated by social science researchers specialising in online consumer interactions. The present work utilises this previously annotated dataset as a benchmark for evaluating machine learning approaches to conflict classification, while contributing new modelling techniques and experimental analysis.

In the original dataset creation process, the researchers adopted a dual-coding methodology to uphold the integrity of the annotation procedure, involving the active participation

of two social science researchers. The initial phase encompassed the first researcher’s deductive identification of incidents involving consumer conflicts. An independent analysis of the data was then undertaken by a second researcher. Both researchers subsequently engaged in a comprehensive assessment of the reliability and applicability of their respective analyses. This deliberation was accompanied by extensive discussions to resolve any divergence in their interpretations. This rigorous and meticulous process culminated in identifying and classifying six distinct categories of conflicts, namely “Teasing”, “Criticism”, “Sarcasm”, “Trolling”, “Harassment”, and “Threats”. These categories span a spectrum of conflict severity and serve as the label space for all models evaluated in this study. The annotations used in this work therefore originate from the previously published dataset, and no additional manual annotation was conducted specifically for the present study. While the dataset itself was introduced in prior work, the contribution of the present study lies in the development and evaluation of novel modelling approaches for conflict classification. Specifically, this paper introduces the ConflictDT reward-based modelling framework, evaluates hierarchical classification strategies, and explores several knowledge distillation methods for improving multi-class conflict detection. The dataset therefore serves as a benchmark for systematically evaluating these modelling approaches rather than representing a newly introduced resource.

Stereotypical hate speech or general toxicity statements may appear across multiple conflict categories depending on their intent and severity. Mild or implicit stereotypes expressed through generalisations may align with the criticism class, whereas derogatory or demeaning stereotypes directed at individuals or groups would typically fall within harassment. In more extreme cases where stereotypes are used to justify intimidation or harm, they may overlap with the threats category. This flexibility reflects the broader aim of the conflict taxonomy to capture behaviours across a spectrum of antagonistic interaction rather than restricting classification to a single linguistic form.

As the six classes used within the task sit on a spectrum of conflict behaviours, theoretical considerations are required when dividing the classes into lesser and more extreme categories of conflict. Teasing, Criticism, and Sarcasm were all selected as less hostile forms of conflict whilst trolling, harassment, and threats were denoted as the more hostile forms of conflict. Sarcasm and teasing can both be considered lighter forms of conflict as there is a significant presence of humour in both, although they still have the potential to cause harm (Wang et al., 2022). Research such as Keltner et al. (2001) supports the similarity of sarcasm and teasing, but reasserts that they are two separate entities. Criticism borders the line between light and severe conflict; although it can be delivered in a constructive manner, it can also be received as a direct insult. Carlson (2016) and Xia (2013) highlight the impact of criticism in different communities within social media. Ultimately, criticism was seen as a socially acceptable form of interaction which, although not always the priority of the perpetrator, can result in harm to the recipient. Alternatively harm is almost exclusively the motive of those conducting trolling, harassment, and threats. These three classes have a clear motivation to cause severe distress to the recipient and therefore are definitively placed within the extreme forms of conflict.

While the netnographic process yielded rich annotations, certain classes—particularly Sarcasm, Threats, and Trolling—were under-represented. To ensure balanced evaluation, we supplemented these categories using external datasets with matching behavioural charac-

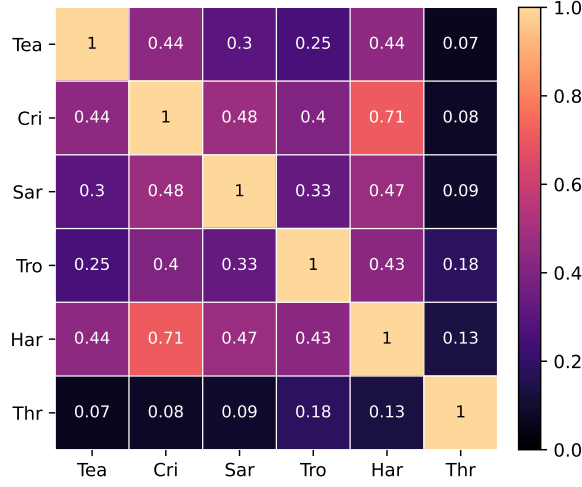


Figure 6: Heatmap showing class cosine similarity.

teristics. The open source datasets chosen were Khodak et al. (2018) for Sarcasm, Wulczyn et al. (2017a) for Threats, and Aggarwal et al. (2020) for Trolling. These external datasets were chosen as the characteristics of the required behaviours within them matched those within this paper. Other researchers (Fortuna et al., 2018; Salminen et al., 2020) investigate this process and find that the practice retains dataset robustness and in some cases can increase task performance. This hybrid construction approach ensures both conceptual integrity and sufficient class representation for training and evaluation. Final conflict dataset statistics are shown in Table 2.

### 3.6 Dataset Characteristics

To examine the robustness of the multi-class conflict dataset utilised in this paper, we conduct an evaluation of its structure, class balance, and semantic distinctiveness. We examine class size distributions, lexical and semantic features, and inter-class similarity using cosine distance metrics, as outlined in the Methodology. The goal is to assess the dataset’s structural integrity and its suitability for training nuanced classifiers. This analytical review helps identify trends and features that could inform model design and evaluation. We also highlight its strengths and limitations to support future research and replication.

A key strength of the conflict dataset is that it encompasses a range of negative behaviours spanning subtle and overt forms of online conflict. As highlighted by Bianchi et al. (2022), there is a scarcity of research on nuanced hate and conflict; this paper addresses that gap by applying a multi-class conflict dataset grounded in established marketing theory and annotated by social science researchers with expertise in online conflict.

Due to the task’s complex and real-life nature, we investigated semantic similarity between the six classes using a heatmap (Figure 6), and observed an average inter-class similarity of 0.32. Whilst similarities between some conflicts are low—e.g., “Threats” and the other

| Class      | Chars | Words | % Stop Words | Number Sentences |
|------------|-------|-------|--------------|------------------|
| Teasing    | 78.5  | 14.6  | 0.31         | 1.6              |
| Criticism  | 232.9 | 42.7  | 0.41         | 2.9              |
| Sarcasm    | 58.9  | 10.7  | 0.32         | 1.1              |
| Trolling   | 105.4 | 18.6  | 0.32         | 1.6              |
| Harassment | 130.2 | 23.9  | 0.35         | 2.2              |
| Threats    | 275.7 | 50.6  | 0.31         | 5.5              |

Table 1: Table showing mean class characteristics within the conflict dataset

classes—there is less distinction between others, such as "Criticism" and "Harassment", which had a high similarity score of 0.71. This overlap poses a challenge for classification, particularly given the dataset’s modest size. Higher similarity between classes results in greater difficulty for classifiers, as the decision boundaries become more ambiguous.

Text brevity is another key factor influencing classification performance. We conducted an analysis of textual features (Table 1), finding that the average character length and word count across all classes were just 147 characters and 26.85 words, respectively. Although prior research has established a correlation between shorter texts and reduced classification performance (Song et al., 2014), the conflict dataset’s brevity accurately reflects the domain’s characteristics and the nature of real-world social media interactions. Nonetheless, developing robust methods to handle short texts and semantic overlap remains essential.

By generating a class similarity matrix, we delve into the dataset’s composition beyond class size, gaining insights that inform both evaluation and the design of our custom reward functions. For instance, we observe that milder conflicts tend to exhibit greater semantic similarity than severe ones. A particularly high level of similarity was found between the "Criticism" and "Harassment" classes. This may reflect behavioural overlap, as previous works (Kim et al., 2022; Jhaver et al., 2018) have investigated how criticism can be perceived as harassment, and how one may evolve into the other depending on context and actor intent. These findings directly inform the reward function design for ConflictDT involving class-level and group-level separation strategies.

Despite its merits, the dataset has limitations. While its size is comparable to other works (Sanh et al., 2019), it remains relatively small compared to large-scale hate speech datasets (Davidson et al., 2017; Vidgen et al., 2021b). However, state-of-the-art deep learning models have demonstrated strong performance on small datasets (Sanh et al., 2019; Caselli et al., 2021), and our experiments confirm that nuanced classification is still feasible. Another drawback is class imbalance: due to the nature of social media discourse, not all conflict types are equally represented. This imbalance must be accounted for in model training and evaluation.

| Lesser Conflicts |            | More Extreme Conflicts |            |
|------------------|------------|------------------------|------------|
| Class            | Datapoints | Class                  | Datapoints |
| Teasing          | 208        | Trolling               | 1089       |
| Criticism        | 698        | Harassment             | 1098       |
| Sarcasm          | 577        | Threats                | 482        |

Table 2: Table showing conflict dataset class sizes

## 4 Implementation and Baselines

### 4.1 Implementation

All models were trained for four epochs using a learning rate of  $2e-5$  and the same Cross Entropy Loss Function. For a classification task, the number of output features is equal to the number of classes. We optimised models using AdamW (Loshchilov and Hutter, 2017). For the baseline models, all hyperparameters were used from the published papers, trained using the same setup and fine-tuned on the same training, validation, and test splits from each dataset. The train, validation, and test splits for all datasets were 80%, 10%, and 10% respectively. For the sentence transformer, we used the all-mpnet-base-v2 (Reimers and Gurevych, 2021)

### 4.2 Metrics and Statistical Tests

To evaluate classifier performance, we employed four key metrics: Accuracy, Precision, Recall, and F1-Score. Given the imbalanced nature of the conflict dataset, both Accuracy and F1-Score were particularly important for assessing model robustness. To further validate the significance of the observed performance differences, we conducted statistical testing using a paired two-tailed t-test. To obtain the performance metrics for each model, we performed 5-fold cross-validation and report the average macro results. When comparing model performance for each dataset we calculated the t-value between the enhanced performance models and the base BERT model. The t-value was then used to calculate a p-value which was evaluated against alphas of  $\alpha = 0.05$  and  $\alpha = 0.10$  to account for borderline effects in small sample comparisons.

### 4.3 Baselines

The following baselines were used:

- BERT: due to proven performance in classification tasks, especially within social media text problems (Wang et al., 2020a; Gupta et al., 2020). For the BERT model, we used the BERT-Base uncased pre-trained model with 12 layers, 12 heads, 768 hidden size, and 110M parameters.
- Flan-T5: A generative language model which can be fine tuned for text classification. Edwards and Camacho-Collados (2024) evaluate Flan-T5 for a multitude of text classification problems, showing strong performance metrics. We use the Flan-T5 base

model with 248M parameters uploaded to the Hugging Face repository by the Google team (Chung et al., 2024; et al., 2023).

- GPT-2 : demonstrated effectiveness in text classification tasks and its use in Decision Transformer work by Minaee et al. (2021); Balkus and Yan (2024); Chen et al. (2021). For the GPT-2 model, we used the default GPT-2 model with 12 layers, 12 heads, 768 hidden size, and 117M parameters.
- HateBERT: (Caselli et al., 2021), focusing on abusive language detection, specifically offensive, abusive and hateful language. HateBERT features intensive pre-training on social media comments before being deployed for fine-tuning domain-specific tasks. For HateBERT we used the default model provided by the authors with 12 layers, 12 heads, 768 hidden size, and 110M parameters.
- DistilBERT: a lightweight variation of the base BERT model, has been proven to be an excellent competitor to the traditional BERT model, with Sanh et al. (2019) stating: "it is possible to reduce the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster". We elected to benchmark DistilBERT for its successful use by Mutanga et al. (2020) within a hate classification task. For the DistilBERT model we used the default DistilBERT model with 6 layers, 768 hidden size, 12 heads, and 66M parameters.

Generative models, Flan-T5 and GPT-2, were also included to assess whether sequence-to-sequence architectures offer advantages in nuanced classification tasks. These baselines provide a diverse set of architectures commonly used in text classification research, including encoder-based models (BERT, DistilBERT, HateBERT), generative transformer models (GPT-2), and instruction-tuned sequence-to-sequence models (Flan-T5). This allows comparison of the proposed approaches across different transformer architectures while maintaining models that are widely used and reproducible in the research community.

The focus of this work is on evaluating modelling strategies for conflict classification rather than improvements driven purely by increasing model size. For this reason, we prioritise widely adopted transformer baselines used in antagonistic and negative behaviour research that allow consistent fine-tuning and comparison across experiments. Larger instruction-tuned LLMs such as LLaMA or GPT-3.5 have recently demonstrated strong performance on a variety of tasks and would therefore represent a natural direction for future evaluation of the proposed methodology.

#### 4.4 Experimental Models

- ConflictDT: A variation of a BERT model with a class-based reward function based on the Decision Transformer Framework. Variations of this model are tested such as various reward functions and a sequential model.
- Hierarchical BERT: A BERT-based hierarchical model which exploits the relationship between lesser and more extreme conflict behaviours. Features a novel loss function defined in the methodology.

- **Knowledge Distillation Models:** Three novel applications of Knowledge Distillation models are used within the paper. Soft Target Distillation leverages softened output distributions from the teacher to guide the student. Feature Distillation using Probabilities introducing a regression head to align the student’s probability vectors with those of the teacher. Feature Distillation using Embeddings aligning the student’s internal hidden states with those of the teacher model.

## 4.5 Datasets

We evaluate all models across three datasets; the conflict dataset introduced in this paper (Breitsohl et al., 2018), (Table 1) with 4,052 texts across "Teasing", "Sarcasm", "Criticism", "Harassment", "Trolling", and "Threat" classes, a subset of the Davidson et al. dataset (Davidson et al., 2017) with texts across "Normal", "Offensive Language", and "Hate Speech" classes, and a subset of the Founta et al. dataset (Founta et al., 2018) with texts across "Abusive", "Hateful", and "Normal" classes. These datasets were selected to span a range of classification tasks both within the hate, negativity, and conflict domain and outside it, enabling evaluation of model generalization across varied task characteristics. Subsets were used for the Davidson et al. and the Founta et al. datasets as the original datasets contained very large numbers of datapoints which would have made training and evaluating models over these datasets prohibitively expensive for this study. We therefore chose to randomly sample subsets which preserve class balance. This allowed us to evaluate on these datasets to investigate model robustness whilst preserving computational efficiency. We reduce the size of each dataset to match the size of our conflict dataset with approximately 4,000 datapoints in each subset, equally distributed across all classes.

## 5 Experiments and Results

### 5.1 Model Performance

#### 5.1.1 EXPERIMENTAL SETUP

This experiment evaluates the performance of ConflictDT and knowledge distillation models against a set of state-of-the-art baselines. We conduct this evaluation across three datasets, our multi-class conflict dataset and two widely used hate speech datasets (Davidson et al. and Founta et al.), to assess generalization and robustness. Several researchers have emphasised the importance of cross-dataset generalisation in social media classification tasks (Fortuna et al., 2021; Barbiero et al., 2020), which is particularly relevant given the modest size of our primary dataset. By benchmarking ConflictDT across multiple domains, we demonstrate its adaptability and validate its comparative performance. This experiment also evaluates the Hierarchical model on the conflict dataset, enabling direct comparison with ConflictDT in a shared classification setting.

#### 5.1.2 RESULTS

Tables 3, 4, and 5 report classification model performance across three datasets (Conflict, Davidson, and Founta), using accuracy, F1, recall, and precision. The hierarchical model cells are empty for Davidson and Founta as only the conflict dataset had a hierarchical

| Model                   | Acc         | F-1           | Rec         | Pre         |
|-------------------------|-------------|---------------|-------------|-------------|
| BERT                    | 0.69        | 0.61          | 0.61        | <i>0.65</i> |
| HateBERT                | 0.55        | 0.52          | 0.52        | 0.52        |
| DistilBERT              | 0.55        | 0.53          | 0.52        | 0.54        |
| GPT-2                   | 0.59        | 0.51          | 0.51        | 0.55        |
| Flan-T5                 | 0.67        | 0.60          | 0.61        | 0.61        |
| Conflict DT             | <i>0.71</i> | <i>0.63*</i>  | <i>0.64</i> | 0.62        |
| Hierarchical BERT       | 0.69        | 0.62          | 0.63        | 0.63        |
| KD Hidden Embedding     | <b>0.72</b> | <b>0.67**</b> | <b>0.67</b> | <b>0.67</b> |
| KD Logits Probabilities | <i>0.71</i> | 0.61          | 0.63        | 0.62        |
| KD Logits Combined Loss | 0.70        | 0.60          | 0.61        | 0.61        |

Table 3: Model performance on Conflict dataset. Best results in bold, second best in italics. The notation \*\* and \* denote statistical significance at p=0.05 and p=0.1 respectively.

| Model                   | Acc         | F-1           | Rec         | Pre         |
|-------------------------|-------------|---------------|-------------|-------------|
| BERT                    | 0.87        | 0.86          | 0.86        | 0.86        |
| HateBERT                | 0.86        | 0.86          | 0.86        | 0.87        |
| DistilBERT              | 0.85        | 0.86          | 0.86        | 0.86        |
| GPT-2                   | <i>0.90</i> | 0.73          | 0.72        | 0.73        |
| Flan-T5                 | 0.87        | 0.87          | 0.86        | <i>0.88</i> |
| Conflict DT             | 0.89        | <i>0.88*</i>  | <i>0.88</i> | <i>0.88</i> |
| Hierarchical BERT       | -           | -             | -           | -           |
| KD Hidden Embedding     | <b>0.91</b> | <b>0.90**</b> | <b>0.90</b> | <b>0.90</b> |
| KD Logits Probabilities | 0.86        | 0.85          | 0.85        | 0.85        |
| KD Logits Combined Loss | 0.86        | 0.85          | 0.85        | 0.85        |

Table 4: Model performance on Davidson et al. dataset. Best results in bold, second best in italics. The notation \*\* and \* denote statistical significance at p=0.05 and p=0.1 respectively.

structure. The results allow us to answer R.Q.1 How effective are the different classification models for improving multi-class classification? How do the different forms of knowledge distillation compare in performance?

Overall, the best performing model was KD Hidden embeddings which consistently outperformed other models in all metrics. Its strong performance suggests that internal representation alignment is particularly effective for capturing nuanced behavioural distinctions, and may offer a more stable signal than output-level supervision. The ConflictDT model although significantly outperformed by KD Hidden Embeddings was the strongest among remaining models. This sets up the discussion in the performance cost analysis (section 5.5). We will now examine the results dataset by dataset to understand how different models compare in specific domains.

| Model                   | Acc         | F-1           | Rec         | Pre         |
|-------------------------|-------------|---------------|-------------|-------------|
| BERT                    | 0.77        | 0.76          | 0.77        | <i>0.78</i> |
| HateBERT                | <i>0.78</i> | 0.73          | 0.73        | 0.73        |
| DistilBERT              | 0.77        | 0.69          | 0.67        | 0.70        |
| GPT-2                   | 0.77        | 0.72          | 0.73        | 0.72        |
| Flan-T5                 | 0.77        | 0.76          | 0.76        | <i>0.77</i> |
| Conflict DT             | 0.77        | <i>0.77</i>   | <i>0.78</i> | <i>0.78</i> |
| Hierarchical BERT       | -           | -             | -           | -           |
| KD Hidden Embedding     | <b>0.83</b> | <b>0.83**</b> | <b>0.83</b> | <b>0.84</b> |
| KD Logits Probabilities | <i>0.78</i> | <i>0.77</i>   | 0.77        | <i>0.78</i> |
| KD Logits Combined Loss | 0.74        | 0.73          | 0.74        | 0.73        |

Table 5: Model performance on Founta et al. dataset. Best results in bold, second best in italics. The notation \*\* and \* denote statistical significance at  $p=0.05$  and  $p=0.1$  respectively.

Within the conflict dataset, the best-performing model was the KD model with hidden state embeddings, which significantly outperformed other models. Whilst the KD Logit Probabilities model improves upon the baseline, it does not do so significantly. It showed modest gains, indicating that output-level supervision can be helpful but may lack the representational depth needed for subtle class boundaries. The KD Combined Loss model reduces performance compared to the baseline BERT model, suggesting that combining soft targets with classification loss—without access to raw logits—may dilute the learning signal or introduce conflicting gradients. These results highlight the importance of distillation design and underscore that not all KD strategies generalise equally across tasks. We also see that ConflictDT ranks second best, outperforming the baseline BERT model by 2% in F1 score and accuracy. This shows that the reward function within ConflictDT helps capture subtle differences. The HierBERT model results in a 1% increase in F1 score but no increase in accuracy; the increase in F1 score is not significant. Although this is not in line with the main experiment in the original HierBERT paper—which showed significant performance increases over the baseline BERT model—it does correlate with the experiments on the two additional tasks conducted in the paper. The additional tasks, like our paper, feature multiple subtasks in both higher-level groups, as opposed to just one in the main task of the HierBERT paper. The results suggest that the HierBERT model struggles when more complex sub-decisions must be made within the groups. Another reason for HierBERT’s performance in this paper’s task is that the difference between the two higher-level groups is less pronounced than in the original paper’s main task. The difference between positive and negative emotions is more distinct than the differences between lesser and more extreme conflict behaviours—especially when some conflict behaviours straddle the border of the split. This idea will be explored more thoroughly in the next experiment, where we evaluate the ability of HierBERT and ConflictDT to differentiate between the two higher-level groups.

On the Davidson et al. dataset (Davidson et al., 2017), the KD hidden state embeddings model performed best whilst ConflictDT also outperformed all other variations of BERT by

2% in F1 score. However, it only marginally outperformed Flan-T5 by 1% in F1 score. GPT-2 outperforms ConflictDT by 1% in accuracy, although this is countered by significantly worse performance in F1 score, where ConflictDT outperforms GPT-2 by 15%.

As can be seen in Table 5, with the exception of the KD hidden state embeddings model, the ConflictDT model with a reward function of distance between all classes outperforms other models in F1 score on the Founta dataset (Founta et al., 2018). We achieve superior performance to BERT in F1 score by 1%, whilst also significantly outperforming GPT-2 by 5%. The best-performing model for accuracy was HateBERT with 0.78, although importantly, ConflictDT outperformed HateBERT in F1 score by 4% and matched all other models in accuracy at 0.77.

To test statistical significance, we performed paired t-tests between the ConflictDT model and the base BERT model. In each test, there were 4 degrees of freedom. For the conflict dataset, the  $t$ -value was 1.53 and the  $p$ -value was 0.08. For the Davidson dataset, the  $t$ -value was 1.62 and the  $p$ -value was 0.07. Therefore, we can say that our ConflictDT model significantly outperforms the base BERT model on the Davidson and Conflict datasets at a threshold of  $\alpha = 0.10$ , but not at  $\alpha = 0.05$ . For the Founta et al. dataset, the t-test did not show a significant  $p$ -value, with a  $t$ -value of 1.13 and a  $p$ -value of 0.15. Although this result was not significant using thresholds of either  $\alpha = 0.05$  or  $\alpha = 0.10$ , the ConflictDT was still tested using 5-fold cross-validation and thus still shows merit in its performance.

Flan-T5 demonstrates strong generalisation across datasets, particularly in precision and recall, and performs competitively with ConflictDT and BERT variants. Its generative architecture may contribute to its robustness in handling diverse text styles. HateBERT, while achieving solid performance on the Founta and Davidson datasets, shows reduced performance on the conflict dataset—likely due to its pretraining on overtly abusive language, which may limit its sensitivity to subtler behaviours. GPT-2’s high accuracy but low F1 score suggests overconfidence in dominant classes, resulting in poor balance across minority labels.

The results across three popular datasets show our ConflictDT model performs well within hate and conflict text classification. Although the best-performing model in the conflict dataset was the KD model with hidden embeddings, the ConflictDT model still ranks second best, outperforming the other two KD models. With regards to the F1 score metric, ConflictDT outperforms the other SOTA models on two datasets whilst achieving marginally better performance on the third. By showing that our model produces robust results, whilst also generalising across datasets, we increase the reliability of our results. This, therefore, mitigates the shortcomings of the conflict dataset, which suffers from class imbalance and data size. In addition, it also shows the adaptability of the reward functions for new data scenarios. Taken together, the comparative results suggest that KD Hidden Embeddings offers the best raw performance, Flan-T5 and HateBERT provide strong domain-specific baselines, and ConflictDT delivers a flexible, interpretable alternative with competitive results and lower computational cost. Finally, although KD models show robust performance, it does come at a higher computational cost. We will evaluate this performance-cost tradeoff later in the paper.

## 5.2 ConflictDT Reward Function Analysis and Sequential Modelling

### 5.2.1 EXPERIMENTAL SETUP

This experiment investigates the role of reward functions and sequential modelling within the ConflictDT architecture, focusing on their impact on classification performance over the conflict dataset. Our approach in selecting individual reward functions involved a thorough analysis of distinct properties within the conflict dataset, specifically addressing existing misclassification patterns and class cross-talk.

We conducted a series of tests with multiple reward functions, examining the impact of each on overall classification performance. This setup showcases the novelty and flexibility of the reward functionality within ConflictDT, which allows the prioritization of different behaviours and aspects. To further analyse the performance of the BERT and ConflictDT models, we also include a breakdown of class-level performance within the conflict dataset.

The reward function was varied while keeping all other parameters constant. We tested five different reward functions:

- i Distance between the text embedding and all classes
- ii Distance between the text embedding and lesser and more extreme forms of conflict
- iii Distance between the "Harassment" class and the text embedding
- iv Distance between the text embedding and all classes with sequential functionality

These reward functions were selected based on our knowledge of the dataset and analysis conducted in section 3.5. The first reward function, "distance between all classes", aimed to prioritise separating all classes within the dataset and could be applied to any classification task. We propose that the extra contextual information will help the model reduce class cross-talk i.e. misclassification. The second reward function, "distance between lesser and extreme conflict groups", aimed to exploit common characteristics within these two groups, as identified during dataset analysis.

The "Harassment"-based reward function was derived from our model's initial results, which indicated that "Harassment" class datapoints were frequently misclassified into other classes and vice versa (see Fig.3). Consequently, we sought to demonstrate the model's ability to prioritise different classification behaviours by designing a reward function to address this trend.

The final two variations of ConflictDT feature sequential modelling. We show the impact of sequential modelling both with and without the reward functionality. This allows us to isolate the contribution of temporal structure and assess whether reward signals enhance or interfere with sequential learning.

We also compare the ability of ConflictDT (using the "lesser vs. extreme" reward function) and the Hierarchical model to differentiate between grouped conflict behaviours. This comparison helps evaluate whether explicit reward guidance or latent structural modelling is more effective in capturing behavioural distinctions.

|                              | Acc         | F-1         | Rec         | Pre         |
|------------------------------|-------------|-------------|-------------|-------------|
| BERT                         | <i>0.69</i> | 0.61        | 0.61        | 0.65        |
| HierBERT                     | <i>0.69</i> | <i>0.62</i> | <i>0.63</i> | 0.63        |
| GPT-2                        | 0.59        | 0.51        | 0.51        | 0.55        |
| Flan-T5                      | 0.67        | 0.60        | 0.61        | 0.61        |
| Dist all Classes             | <b>0.71</b> | <b>0.63</b> | <b>0.64</b> | 0.62        |
| Dist Less & Gtr Class Groups | 0.67        | 0.59        | 0.62        | <b>0.75</b> |
| Dist "Harassment"            | 0.68        | 0.61        | <i>0.63</i> | 0.67        |
| Sequential no Reward         | 0.68        | 0.60        | 0.60        | <i>0.74</i> |
| Sequential w/ Reward         | <i>0.69</i> | <i>0.62</i> | 0.62        | 0.64        |

Table 6: The effects of reward functions in the conflictDT model. Best results in bold, second best in italics.

### 5.2.2 RESULTS

Table 6 shows the effects of different reward functions within the ConflictDT model, which produced a variety of classification performance scores. Here, we explore the potential of various reward functions in directing the classifier toward correct decisions. All reward-based variants of ConflictDT show comparable performance to the BERT baseline, with some outperforming it in specific metrics.

The ConflictDT model with a reward based on distances to all classes outperforms BERT by 4% in F1-score and 2% in accuracy. This variant achieved the best overall performance, demonstrating the effectiveness of general reward signals in guiding model decisions. As in section 5.1, BERT-based models outperform GPT-2 across all metrics.

Rewards based on the distance between lesser and more extreme conflict groups, and between "Harassment" and other classes, are slightly outperformed by BERT in overall accuracy. The "Harassment" reward matches BERT in F1-score, while the group-based reward sees a decrease of 0.02 in F1. Although these variants do not yield overall performance gains, they demonstrate the model's ability to influence behaviour at the class level.

Figure 8 shows individual class performance when using the Harassment reward. While overall metrics remain similar to BERT, the frequency at which other classes are misclassified as Harassment drops. Similarly, the group-based reward does not improve overall performance, but reduces the number of "Lesser" class datapoints misclassified into the "Greater" group. However, it also slightly increases the number of "Greater" datapoints misclassified as "Lesser", indicating a trade-off in directional control. These results show that ConflictDT can be used to influence model behaviour with respect to individual classes or structural properties of the classification task.

Examining BERT and ConflictDT classification heatmaps over the six-class conflict dataset (Figure 7), we observe that ConflictDT achieves higher performance in Teasing, Criticism, and Trolling; equal performance in Sarcasm; and lower performance in Harassment. Additionally, ConflictDT reduces the frequency of other classes being misclassified as Harassment, while Harassment is more frequently misclassified into Sarcasm and Trolling.

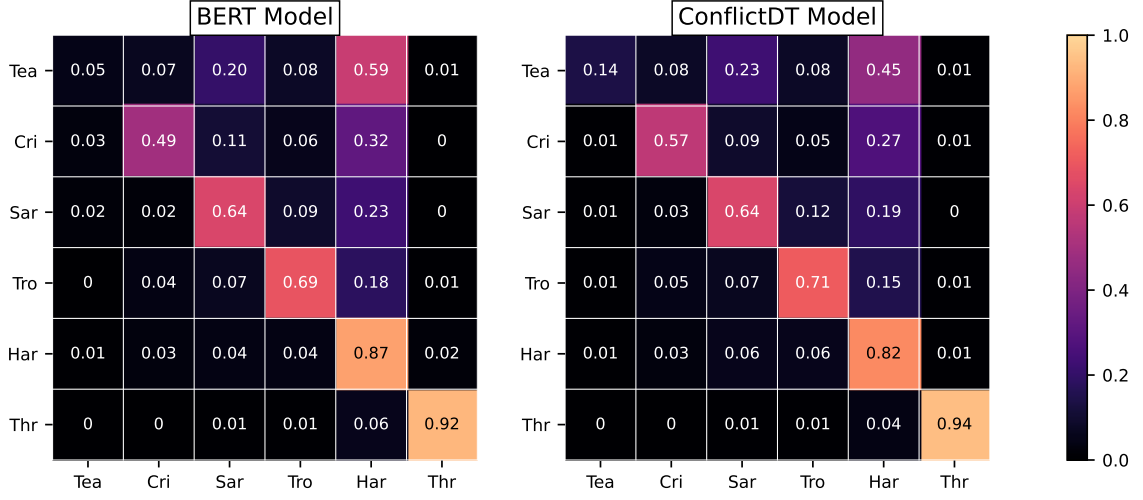


Figure 7: BERT and ConflictDT class performance on the conflict dataset.

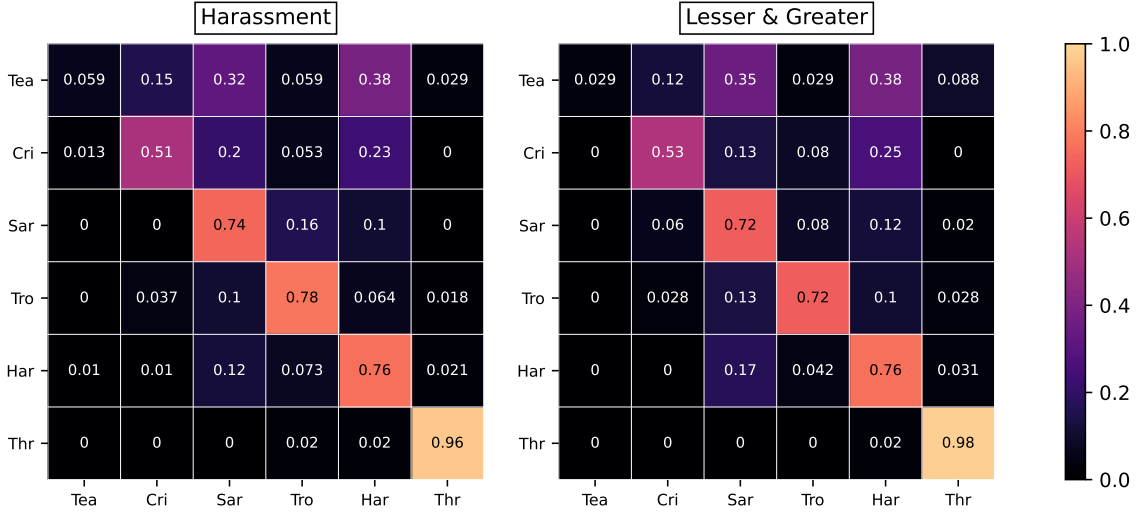


Figure 8: ConflictDT class performance with Harassment and Lesser-Greater rewards on the conflict dataset.

Considering the earlier dataset analysis in section 3.5, this outcome was unexpected. We had theorised that promoting distance between the two class groups would reduce misclassification between them. The observed decrease in performance may be due to ambiguity within each group or insufficient semantic separation between them.

Both sequential versions of the ConflictDT model fail to make significant improvements over the base model, with or without rewards. We theorise that this may be due to the short text length within the conflict dataset, quantified in Table 1. The additional data gained at each timestep may simply not be sufficient to aid the model, and may instead

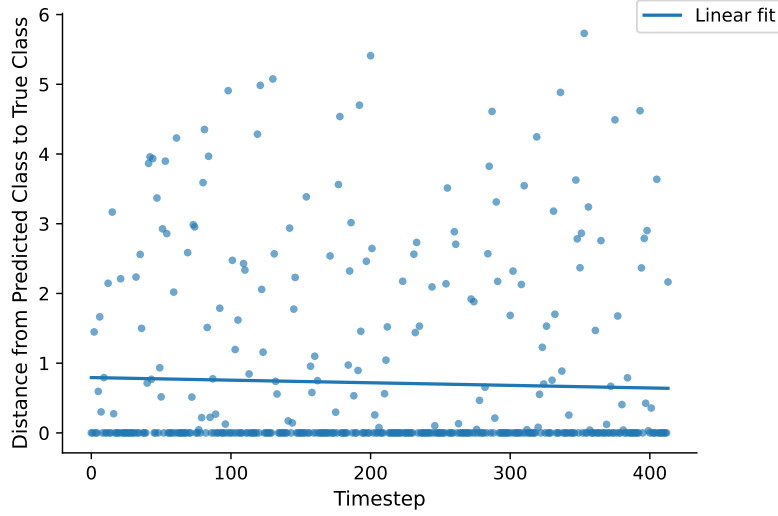


Figure 9: Graph showing the distance between the logits of the predicted class and the logits of the true class changing over timesteps.

introduce noise. To investigate this, we measured how the distance between the logits of the predicted class and the logits of the true class changed over time. If the sequential model worked as intended, this distance would converge toward zero as timesteps increase.

For all datapoints in the test set, we calculated the average distance between the predicted and true class logits at each timestep. Figure 9 shows that the average distance decreases slightly over time. We further calculated the Pearson correlation coefficient between distance and timestep, resulting in a coefficient of  $-0.05$  and a p-value of  $0.34$ . This suggests a weak negative correlation that is not statistically significant, indicating that sequential modelling does not meaningfully improve prediction in this context.

| Model               | Acc  | F1   | Rec  | Pre  |
|---------------------|------|------|------|------|
| HierBERT            | 0.81 | 0.79 | 0.78 | 0.80 |
| ConflictDT Less-Gtr | 0.79 | 0.77 | 0.77 | 0.79 |

Table 7: Table evaluating the binary classification between the lesser and extreme conflict behaviour groups.

Table 7 presents binary classification performance between the lesser and extreme conflict behaviour groups. The HierBERT model outperforms ConflictDT with the group-based reward in both accuracy and F1-score. We can summarise that HierBERT is effective at distinguishing between the two groups, but ultimately does not improve classification at the second step when classifying into individual classes. This second decision is constrained by the initial group assignment—meaning that the 19% of datapoints misclassified at the first

step are automatically incorrect at the second. If the two conflict class groups were more distinct, methods that exploit these differences might be more successful.

In conclusion conflictDT with the reward of distance to all classes is the best performing variant. Targeted rewards such as harassment and lesser/greater distances show class level control but have limited increases in their overall performance. Sequential modelling shows no real gains, likely due to dataset characteristic constraints. A hierarchical model within the complex conflict class is effective at group separations but is limited by ambiguity within the groups. These findings highlight the flexibility of ConflictDT’s reward design, whilst also providing insights into alternative models.

### 5.3 Empirical Analysis of Distillation Weight

#### 5.3.1 EXPERIMENTAL SETUP

This experiment evaluates the impact of distillation loss weighting on model performance across the three knowledge distillation strategies: Soft Target Distillation, Feature Distillation using Probabilities, and Feature Distillation using Hidden State Embeddings. We incrementally increased the weighting of the distillation loss relative to the classification loss, ranging from 1 to 10 in unit steps. This allows us to assess how sensitive each distillation method is to supervision strength and whether optimal weighting varies by strategy. As the raw distillation loss varies across models, using a uniform weighting scheme is unsuitable for direct comparison of absolute values. Instead, we focused on relative performance trends and identify optimal weighting ranges for each approach.

#### 5.3.2 RESULTS

The empirical analysis in Table 5.3.2 clearly shows that for each KD method, there exists an optimal distillation weight. For all three models, performance metrics decline both below and above this optimal point. For KD Logit Probabilities, the optimal distillation weight is 4.0; for KD Hidden Embeddings, it is 7.0; and for KD Soft Combined Loss, the optimal weight is 3.0, although 1.0 is a close second.

Interestingly, KD Hidden Embeddings is the best-performing model overall, KD Logit Probabilities ranks second, and KD Soft Combined Loss performs worst. This suggests a relationship between distillation weight and model effectiveness: more effective distillation strategies tend to benefit from higher weighting.

For the conflict classification task, the traditional classification loss was approximately 10 times larger than the distillation losses. Therefore, we explored a distillation weight range from 1.0 to 10.0. Had the two losses been of similar magnitude, a range from 0.1 to 1.0 may have been more appropriate. This highlights an important consideration: distillation weight—or any hyperparameter—is task-dependent, and fine-tuning is essential for optimal performance.

### 5.4 Thematic Analysis of Misclassifications

#### 5.4.1 EXPERIMENTAL SETUP

This experiment conducts a qualitative analysis of ConflictDT classifier outputs on the six-class conflict dataset, focusing on patterns of misclassification. To complement our

| Wt. | Acc                    | F1          | Pre         | Rec         | Acc                   | F1          | Pre         | Rec         | Acc               | F1          | Pre         | Rec         |
|-----|------------------------|-------------|-------------|-------------|-----------------------|-------------|-------------|-------------|-------------------|-------------|-------------|-------------|
|     | KD Logit Probabilities |             |             |             | Feature Hidden States |             |             |             | Soft Distillation |             |             |             |
| 1   | 0.67                   | 0.58        | 0.57        | 0.58        | <b>0.72</b>           | 0.64        | 0.64        | 0.65        | 0.67              | <b>0.60</b> | <b>0.61</b> | <b>0.61</b> |
| 2   | 0.67                   | 0.60        | 0.60        | 0.59        | 0.69                  | 0.63        | 0.64        | 0.63        | 0.66              | 0.59        | 0.58        | 0.59        |
| 3   | 0.68                   | <b>0.61</b> | 0.61        | 0.61        | 0.70                  | 0.66        | <b>0.70</b> | 0.65        | <b>0.70</b>       | <b>0.60</b> | <b>0.61</b> | <b>0.61</b> |
| 4   | <b>0.71</b>            | <b>0.61</b> | <b>0.62</b> | <b>0.63</b> | 0.70                  | 0.64        | 0.66        | 0.64        | 0.67              | 0.59        | 0.60        | 0.59        |
| 5   | 0.70                   | 0.59        | 0.58        | 0.60        | 0.71                  | 0.66        | 0.66        | 0.67        | 0.63              | 0.58        | 0.58        | 0.59        |
| 6   | 0.67                   | 0.56        | 0.56        | 0.58        | 0.71                  | 0.64        | 0.67        | 0.64        | 0.64              | 0.59        | 0.60        | 0.59        |
| 7   | 0.65                   | 0.56        | 0.56        | 0.58        | 0.71                  | <b>0.67</b> | <b>0.67</b> | <b>0.68</b> | 0.63              | 0.55        | 0.56        | 0.56        |
| 8   | 0.65                   | 0.57        | 0.56        | 0.58        | 0.69                  | 0.62        | 0.64        | 0.63        | 0.59              | 0.52        | 0.52        | 0.51        |
| 9   | 0.63                   | 0.55        | 0.54        | 0.57        | 0.68                  | 0.59        | 0.73        | 0.60        | 0.58              | 0.53        | 0.54        | 0.52        |
| 10  | 0.62                   | 0.55        | 0.54        | 0.57        | 0.69                  | 0.60        | 0.59        | 0.62        | 0.56              | 0.53        | 0.53        | 0.53        |

Table 8: Macro Performance Across Distillation Methods and Weights

quantitative evaluation, we apply thematic analysis—a widely recognised method in the social sciences for interpreting human communication—formally defined by Braun and Clarke (2006). Thematic analysis enables us to identify recurring themes and behavioural patterns within misclassified online comments. We follow the six-phase framework outlined by Braun and Clarke (2021): (1) Familiarization with the data, (2) Generation of initial codes, (3) Searching for themes, (4) Reviewing themes, (5) Defining and naming themes, and (6) Producing the report. One coder annotated the data, followed by a second coder who reviewed the identified codes, patterns, and themes. The two researchers then engaged in discussion to resolve discrepancies and finalise the thematic structure. The final analysis includes definitions, descriptions, and representative sample comments for each theme, enabling replication and interpretability. This approach provides insight into how model errors may reflect deeper communicative ambiguities or social dynamics.

#### 5.4.2 RESULTS

The thematic analysis produced three main misclassification themes, which we refer to as Linguistic Fluidity, Context Dependency, and Humour Ambiguity. These themes—alongside their definitions, descriptions, examples, and frequency—are presented in Table 9.

The theme of Linguistic Fluidity encompasses the blurred boundaries between class behaviours. Although datapoints tend to exhibit a dominant behaviour, they often contain aspects of multiple class behaviours. This ambiguity complicates both human and machine classification. For example, Jhaver et al. (2018) and Kim et al. (2022) identify crossover between Criticism and Harassment, discussing how Criticism can evolve into Harassment and how the perceived identity of a behaviour is often subjective. This theme also emerges in hate speech terminology, where Fortuna et al. (2020) highlight inconsistencies across research papers that lead to ambiguity between behaviour classes and misinterpretation of labels.

The second theme, Context Dependency, relates to datapoints where correct classification depends on conversational context that is unavailable to the model. While individual social media comments may contain overt negative behaviours, they are often embedded within larger threads or chains of interaction. Human annotators conducting netnographies

| Theme                      | Definition                                                                                                                      | Description                                                                                                                                                                                                                                                                                                                                                                                                                             | Examples                                                                                                                                                                                                                                                                                                                                              |
|----------------------------|---------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Linguistic Fluidity</b> | A misclassification that occurs due to the lack of definitional boundaries that are inherent to the interpretation of language. | A commonly known phenomenon in linguistics is that of multiple meanings to the same sentence, where meaning interpretation depends on a multitude of unpredictable factors(e.g. one's mood, need for politeness etc; ) Classes are not always clear cut and often have fluid boundaries. Datapoints can contain behaviour which could belong to more than one class. Therefore neither classifier or human can get it totally accurate. | " Quickly! Let's spend our precious weekend time arguing about fast food on the Internet!"<br><br>"everyone in every country has a right to their own opinion. I'm not sure why it matters what country, it's their opinions and their beliefs. No one is asking you to agree but respect the fact they are entitled to their opinions and beliefs. " |
| <b>Context Dependency</b>  | A misclassification that occurs when the meaning of a message changes when taken out of context.                                | Context dependency is a common issue identified by language researchers when studying offensive/impolite communication. The type of behaviour exhibited in a datapoint depends on the context of the prior comments. When classifiers are utilising single comment datapoints this context is not present, as opposed to the human coders who had access to the whole conversation when conducting netnographies.                       | "Picture that says: it's a joke not a dick, don't take it so hard."<br><br>"If Andy Dwyer wasn't likable."                                                                                                                                                                                                                                            |
| <b>Humour Ambiguity</b>    | A misclassification that occurs when a message fails to convey that it was meant in humour and/or was good vs bad-natured.      | Linguists have long recognised the lack of clarity inherent to humour as a quality on which humour often relies. Humour has been recognised as a particularly challenging area of NLP. Humour can be taken two ways and is subjective meaning the dominant type of humour behaviour is often ambiguous within datapoints.                                                                                                               | "The Not So Fresh Princess of Stale Air."<br><br>"Keep checking on Arsenal every champions league game mate ! Mu won't be there! Arsenal are here to entertain! unlike your confused club!"                                                                                                                                                           |

Table 9: Table showing the themes identified in one test set's misclassifications

| Model             | Runtime<br>(seconds) | Carbon Footprint<br>(g CO <sub>2</sub> e) | Energy Consumption<br>(Wh) |
|-------------------|----------------------|-------------------------------------------|----------------------------|
| <b>ConflictDT</b> | 799.60               | 2.19                                      | 23.01                      |
| <b>Normal</b>     | 761.92               | 2.06                                      | 21.65                      |
| <b>Sequential</b> | 2384.98              | 6.46                                      | 67.76                      |
| <b>LLM Probs</b>  | 8451.17              | 22.96                                     | 241.16                     |
| <b>LLM Hidden</b> | 1115.31              | 3.03                                      | 31.84                      |
| <b>LLM Soft</b>   | 8445.70              | 22.95                                     | 240.99                     |

Table 10: Table showing the computational cost of training classification models.

had access to full conversations, whereas classifiers typically operate on isolated datapoints. This lack of context leads to misclassification when the behaviour is contingent on prior or surrounding discourse. Many researchers have addressed this issue by developing models that incorporate conversational context into classification tasks (Qian et al., 2018; Dahiya et al., 2021; Sahnan et al., 2021; Chakraborty and Masud, 2022).

The final theme, Humour Ambiguity, reflects the difficulty NLP models face in identifying humour, particularly when it is subtle, culturally specific, or reliant on shared knowledge. Humour is inherently subjective and often ambiguous, making it a persistent challenge in NLP. For example, the first humour ambiguity datapoint in Table 9, “The Not So Fresh Princess of Stale Air,” was labelled as “Trolling” but misclassified as “Harassment”. The model failed to detect the humorous intent and cultural reference to a 1990s sitcom, which a human reader would likely interpret as playful rather than hostile. The second example illustrates how teasing between football fans relies on shared knowledge of club rivalries and fan culture. Without this contextual grounding, the model struggles to distinguish banter from aggression.

## 5.5 Performance-Cost Analysis

### 5.5.1 EXPERIMENTAL SETUP

This experiment evaluates the trade-off between classification performance and training cost across six models: ConflictDT, Sequential ConflictDT, KD Embedding, KD Probabilities, KD Soft Loss, and a baseline BERT model. We record the time taken to execute training loops for each model and use this as a proxy for computational expense. To extend this analysis beyond runtime, we apply the Green Algorithm’s carbon cost estimator introduced by Lannelongue et al. (2021). This tool estimates the carbon footprint of any computational task by taking into account GPU type and count, memory usage, runtime, and platform. Using this framework, we present a novel accounting of the environmental cost of training classification models, enabling comparison not only in terms of accuracy but also environmental impact.

### 5.5.2 RESULTS

The results shown in Table 10 illustrate the computational costs of the six evaluated models. Comparing ConflictDT with the BERT model, we observe slight increases of 4.95% in

runtime, 6.31% in carbon footprint, and 6.28% in energy consumption. The Sequential model, as expected, incurs significantly higher training costs, with increases of 314% in runtime, 313% in carbon footprint, and 298% in energy consumption.

The introduction of LLM-based KD methods results in a substantial increase in computational cost. Both KD Logit Probabilities and KD Soft Combined Loss models exhibit runtimes and carbon footprints far exceeding all other models. Given their relatively modest performance gains, these methods appear unsuitable for the conflict classification task. KD Hidden Embeddings also incurs higher runtime, but the increase is more modest and presents a realistic option.

Evaluating performance against computational cost, we find that the Sequential models—despite costing roughly three times more to run—offer no improvement in classification performance. ConflictDT and KD Hidden Embeddings present a more nuanced trade-off. ConflictDT increases F1-score by 2% with only a 4.95% increase in runtime, while KD Hidden Embeddings improves F1-score by 6% but increases runtime by 46.4%. In other words, ConflictDT is more efficient in terms of performance gain per unit cost, whereas KD Hidden Embeddings delivers the highest raw performance.

In conclusion, the Normal classification model offers the best value per performance. However, in scenarios where performance is paramount, KD Hidden Embeddings should be preferred. If deploying a large LLM for distillation is not feasible, but improved performance is still required, ConflictDT offers a practical middle ground—delivering gains at only a modest increase in cost.

## 6 Discussion of Results

Here we present a summary of the key findings in our results in relation to the research questions defined at the start of this paper.

### 6.1 Effectiveness of Classification Models

We successfully answer RQ1. by showing that some classification models are highly effective and significantly outperform the baseline. Notably, knowledge distillation with hidden embeddings and ConflictDT were the top two models respectively. Not all knowledge distillation methods were equal in performance, with knowledge distillation with logits and combined loss underperforming the baseline. These increases in performance are vital for real world application, especially within a high stakes conflict classification scenario,

### 6.2 ConflictDT Reward Performance

We confirm RQ2. by showing that the reward function within ConflictDT can be used to influence model behaviour and increase classification performance. Not all proposed rewards increased performance, but even some of those still achieved some of their goals, for example "Lesser and Greater" reward decreases the number of datapoints misclassified between the groups even though it doesn't decrease the number misclassified overall significantly. This ability to influence model behaviour is highly promising and positions ConflictDT as a novel method within text classification.

### 6.3 Hierarchical Framework Performance

We were able to show that the hierarchical model outperformed the ConflictDT model which focused on the parent group but both models were successful at differentiating between the parent groups (Shown in Table 7). This answers RQ3, showing that both methods capture the latent structure within the dataset. The hierarchical model does so by strictly differentiating between the parent groups whilst ConflictDT does so via its more subtle reward based approach. This experiment and RQ3 reinforce ConflictDT’s ability to exploit relationships between classes alongside individual class characteristics to improve performance.

### 6.4 Knowledge Distillation Hyperparameters

As mentioned we were able to show that knowledge distillation with hidden state embeddings and knowledge distillation with probability embeddings improved classification performance over the baseline. We were further able to show through an empirical study that distillation hyperparameters have a significant effect on model performance. Not all distillation methods had the same distillation weight for optimum performance indicating that tuning per method and task is required. Results also show diminishing returns when tuning hyperparameters too aggressively. These results provide a clear and satisfactory answer to RQ5.

### 6.5 Model Misclassification Patterns

This experiment resulted in some interesting findings through the interdisciplinary methods which answered RQ4. The thematic analysis revealed three themes of misclassification; linguistic fluidity, context dependency, and humour ambiguity. These model misclassification trends link to existing social science theory, showing that models mirror the difficulties experienced by humans when interpreting nuanced online conflict behaviours.

### 6.6 Model Performance-Cost Tradeoffs

This experiment succinctly answered RQ.6 by showing the performance cost tradeoff using three key examples. Both ConflictDT and KD improve over the baseline model but both require extra computation. ConflictDT improves performance over the baseline without requiring large teacher models, making it cost effective. KD, whilst clearly offering the best performance, requires expensive teacher training. This creates questions surrounding cost, efficiency, and scalability within real world deployment.

### 6.7 Implications for Moderation and Responsible AI

The findings of this study have implications for automated moderation systems deployed in online platforms. Most current moderation pipelines focus on detecting overt forms of abuse such as harassment or explicit hate speech. Our results demonstrate that models can also learn to distinguish between more subtle conflict behaviours, such as teasing, sarcasm, and criticism. Detecting these earlier-stage behaviours may enable platforms to identify potentially harmful interactions before they escalate into more severe hostility.

At the same time, the misclassification analysis highlights the challenges inherent in automating nuanced behavioural interpretation. Linguistic fluidity, humour, and context dependency mean that even advanced models struggle to interpret intent in some cases.

This reinforces the importance of deploying automated classifiers as assistive tools rather than fully autonomous moderation systems. Human oversight and contextual judgement remain essential components of responsible content moderation.

This work also contributes to ongoing interdisciplinary discussions between computer science and social science regarding the measurement and interpretation of online conflict behaviours. By combining machine learning evaluation with qualitative thematic analysis, the study demonstrates how computational models can both benefit from and contribute to broader theoretical understandings of online communication.

## 7 Threats to Validity and Limitations

### 7.1 Dataset and External Validity

While evaluation across three datasets improves the robustness of our findings, several limitations remain. The conflict dataset introduced in this work is relatively modest in size compared to many contemporary NLP benchmarks and exclusively originates within online brand community discussions. Although the Davidson et al. and Founta et al. datasets were included to assess generalisation, the results may not fully transfer to other social media platforms, domains, or communication settings. Furthermore, the datasets used in this study are primarily English-language datasets and therefore do not capture linguistic and cultural variation that may influence the expression and interpretation of conflict behaviours. Future work should evaluate the proposed methods on larger, multilingual, and cross-platform datasets

### 7.2 Construct and Annotation Validity

The classification of online conflict behaviours is inherently subjective. Behaviours such as teasing, sarcasm, criticism, and harassment often exhibit overlapping characteristics and may be interpreted differently depending on cultural norms, conversational context, and annotator perspectives. Although the original dataset employed a dual-coding methodology to improve annotation reliability, some degree of ambiguity remains unavoidable. This limitation is further supported by the thematic analysis, which identified linguistic fluidity, context dependency, and humour ambiguity as recurring sources of misclassification. Consequently, the reported results should be interpreted as modelling socially constructed behavioural categories rather than objectively separable classes and any automated conflict detection should be interpreted cautiously and deployed with appropriate human oversight.

### 7.3 Experimental and Evaluation Validity

The study evaluates models using accuracy, precision, recall, and macro F1-score, which provide a broad view of classification performance across imbalanced classes. However, additional evaluation perspectives, such as calibration analysis, robustness testing, fairness assessment, and uncertainty estimation, were outside the scope of the present work. Furthermore, while the selected baselines represent widely adopted transformer architectures within text classification research, the evaluation does not include comparisons against more recent large instruction-tuned language models. Such comparisons would provide a stronger benchmark for assessing the relative effectiveness of the proposed approaches.

## 7.4 Reproducibility and Data Availability

To support reproducibility, all model architectures, training procedures, and hyperparameters are reported within the paper, and the implementations of ConflictDT, the hierarchical models, and the knowledge distillation approaches are publicly available. However, the conflict dataset contains sensitive social media content and cannot be distributed without restrictions. Access to the dataset is therefore provided on request for research purposes, subject to appropriate ethical approval and data-use considerations. While this may limit immediate replication, it represents a balance between research transparency and responsible handling of potentially harmful content.

## 8 Future Work

Future work should further evaluate the proposed approaches against stronger and more diverse baselines, including recent instruction-tuned and open-source large language models. Additional evaluation measures, such as calibration, robustness, uncertainty estimation, and fairness metrics, would provide a more comprehensive assessment of model behaviour in high-stakes moderation scenarios. Future studies should also investigate multilingual and cross-platform settings to better understand the generalisability of the proposed methods. To support continued research in this area, future work will focus on improving reproducibility through the public release of model implementations, training configurations, and evaluation pipelines, while maintaining controlled access to sensitive datasets.

Additionally, there is room for further exploration of reward functions. Our experiments focused on the distance between class text embeddings, but future reward functions could encompass other aspects, such as intra-class datapoint similarity, emotion and sentiment alignment, or behavioural clustering based on user intent. These alternatives may offer more targeted control over model behaviour and improve performance in ambiguous cases.

Research could also focus on utilising sequential modelling within text classification. The sequential variants of the ConflictDT architecture did not produce significant performance improvements in our experiments. One likely explanation is the short length of the social media posts within the datasets used in this study, which limits the amount of contextual information available for sequential modelling approaches to exploit. The Decision Transformer work by Chen et al. (2021) showed sequential modelling to achieve great success in other IR tasks. Given the sequential nature of text, the Decision Transformer framework could be exploited further, perhaps for long text forms or considering alternative representations of states within the text, such as individual words, different groups of words, or even paragraphs. Exploring hierarchical or multi-resolution state representations may help overcome the limitations observed in short-text scenarios.

Future work could also investigate more advanced knowledge distillation strategies tailored to conflict classification. For example, multi-teacher distillation could be used to transfer complementary behavioural insights from models trained on different domains (e.g. sarcasm detection, hate speech, trolling). Contrastive distillation techniques may help sharpen class boundaries by encouraging separation between semantically similar but behaviourally distinct classes. Additionally, task-adaptive distillation schedules—where the distillation weight varies dynamically based on class difficulty or model confidence—could offer more

flexible and context-aware supervision. Exploring multilingual distillation could also be valuable, particularly for cross-cultural conflict detection in global social media platforms.

Although we provide a qualitative analysis of our model’s misclassifications using a popular and well-recognised social science technique, there is potential to perform a quantitative explainable AI (XAI) experiment. This would aid in understanding the role that the reward functionality plays within the model, and could help identify which features or signals are most influential in shaping predictions.

Finally, a case study investigating the use of this model within a simulated real-life scenario could be conducted, exploring its feasibility as a tool for use within social media platforms. Such a study would demonstrate the practical impact of a model capable of identifying a range of behaviours, while also highlighting responsible AI practices and ethical deployment considerations.

## 9 Conclusion

In this study, we introduce new modelling approaches for multi-class conflict classification and systematically compare three strategies—ConflictDT, hierarchical classification, and knowledge distillation—to assess their effectiveness. Our research findings indicate that the novel reward function within ConflictDT can successfully guide classification model behaviour and improve performance without relying on large teacher models. The hierarchical model was shown to be able to leverage the latent structure within the conflict groupings, although did not improve overall performance. Through the novel application of Knowledge distillation within the text conflict classification task we were able to replicate prior works within different domains to show that Knowledge distillation is a very effective method to improve classification performance. Together, this extensive study of classification models and techniques beyond baseline models advances conflict and text classification, highlighting the importance of analysing the full spectrum of conflict behaviours.

We also extend previous research by applying three knowledge distillation methods and evaluating distillation weights, showing that the performance of different knowledge distillation methods and hyperparameters are not uniform. To evaluate the effectiveness of the classification models we conducted a performance-cost evaluation, showing that although Knowledge distillation with hidden embeddings was the best performing model, ConflictDT achieved the most efficient performance increase. This underscores the importance of balancing performance with scalability. Finally, we conduct a qualitative analysis of model misclassifications, employing thematic analysis to identify trends and patterns within misclassifications. Themes such as linguistic fluidity, context dependency, and humour ambiguity reveal the limitations of current classifiers and highlight the need for socially informed modelling strategies. Overall, our research suggests the classification system and custom dataset introduced in this paper can serve as the foundation for further exploration into social media conflict.

## Acknowledgments and Disclosure of Funding

Oliver Warke was supported by an EPSRC-funded PhD studentship at the School of Computing Science, University of Glasgow. No other additional external funding was received for this research.

## References

- Elias Aboujaoude, Matthew W Savage, Vladan Starcevic, and Wael O Salame. Cyberbullying: Review of an old problem gone viral. *Journal of adolescent health*, 57(1):10–18, 2015.
- Karmanya Aggarwal, Pakhi Bamdev, Debanjan Mahata, Rajiv Ratn Shah, Ponnurangam Kumaraguru, et al. Trawling for trolling: A dataset. *arXiv:2008.00525*, 2020.
- Waseem Akram and Rekesh Kumar. A study on positive and negative effects of social media on society. *International journal of computer sciences and engineering*, 5(10):351–354, 2017.
- Davey Alba and Kurt Wagner. Elon Musk cuts more Twitter staff overseeing content moderation, Jan 2023. URL <https://www.bloomberg.com/news/articles/2023-01-07/elon-musk-cuts-more-twitter-staff-overseeing-content-moderation>.
- Mohsan Ali, Mehdi Hassan, Kashif Kifayat, Jin Young Kim, Saqib Hakak, and Muhammad Khurram Khan. Social media content classification and community detection using deep learning and graph analytics. *Technological Forecasting and Social Change*, 188: 122252, 2023.
- Fatimah Alkomah and Xiaogang Ma. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273, 2022.
- Brooke Auxier and Monica Anderson. Social media use in 2021. *Pew Research Center*, 1: 1–4, 2021.
- Salvador V Balkus and Donghui Yan. Improving short text classification with augmented data using GPT-3. *Natural Language Engineering*, 30(5):943–972, 2024.
- Pietro Barbiero, Giovanni Squillero, and Alberto Tonda. Modeling generalization in machine learning: A methodological and computational study. *arXiv:2006.15680*, 2020.
- Chloe Berryman, Christopher J Ferguson, and Charles Negy. Social media use and mental health among young adults. *Psychiatric quarterly*, 89:307–314, 2018.
- Federico Bianchi, Stefanie Hills, Patricia Rossini, Dirk Hovy, Rebekah Tromble, and Nava Tintarev. “It’s not just hate”: A multi-dimensional perspective on detecting harmful speech online. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 8093–8099, 2022.
- Layla Borooun, Babak Abedin, and Eila Erfani. The dark side of using online social networks: a review of individuals’ negative experiences. *Journal of Global Information Management (JGIM)*, 29(6):1–21, 2021.
- Mondher Bouazizi and Tomoaki Ohtsuki. Multi-class sentiment analysis on Twitter: Classification performance and challenges. *Big Data Mining and Analytics*, 2(3):181–194, 2019.
- Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

- Virginia Braun and Victoria Clarke. Can I use TA? Should I use TA? Should I *not* use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and psychotherapy research*, 21(1):37–47, 2021.
- Jan Breitsohl, Holger Roschk, and Christina Feyertag. Consumer brand bullying behaviour in online communities of service firms. *Service Business Development: Band 2. Methoden–Erlösmodelle–Marketinginstrumente*, pages 289–312, 2018.
- Thomas Brewster. Musk’s X fired 80% of engineers working on trust and safety, australian government says, Aug 2024. URL <https://www.forbes.com/sites/thomasbrewster/2024/01/10/elon-musk-fired-80-per-cent-of-twitter-x-engineers-working-on-trust-and-safety/>.
- Stefano Brogi. Online brand communities: a literature review. *Procedia-Social and Behavioral Sciences*, 109:385–389, 2014.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Taina Bucher, Anne Helmond, et al. The affordances of social media platforms. *The SAGE handbook of social media*, 1:233–253, 2018.
- Matt Carlson. Embedded links, embedded meanings: Social media commentary and news sharing as mundane media criticism. *Journalism studies*, 17(7):915–924, 2016.
- Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, 2021.
- Tanmoy Chakraborty and Sarah Masud. Nipping in the bud: detection, diffusion and mitigation of hate speech on social media. *ACM SIGWEB Newsletter* (Winter):1–9, 2022.
- Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.
- Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Antisocial behavior in online discussion communities. In *Proceedings of the international aaai conference on web and social media*, volume 9, pages 61–70, 2015.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.
- Snehl Dahiya, Shalini Sharma, Dhruv Sahnan, Vasu Goel, Emilie Chouzenoux, Víctor Elvira, Angshul Majumdar, Anil Bandhakavi, and Tanmoy Chakraborty. Would your tweet invoke hate on the fly? Forecasting hate intensity of reply threads on Twitter. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2732–2742, 2021.

- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515, 2017.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
- Ashwin Geet d’Sa, Irina Illina, and Dominique Fohr. Classification of hate speech using deep neural networks. *Revue d’Information Scientifique & Technique*, 25(01), 2020.
- Clare Duffy. Meta gets rid of fact checkers and says it will reduce ‘censorship’, CNN business, Jan 2025. URL <https://edition.cnn.com/2025/01/07/tech/meta-censorship-moderation/index.html>.
- Aleksandra Edwards and Jose Camacho-Collados. Language models for text classification: Is in-context learning enough? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10058–10072, 2024.
- F Elsafoury. Cyberbullying datasets. *Mendeley*. URL <https://data.mendeley.com/datasets/jf4pzyvnpj/1>, 2020.
- Hyung Chung et al. Hugging face, Google FLAN-T5 base, 2023. URL <https://huggingface.co/google/flan-t5-base>.
- Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.
- Paula Fortuna, José Ferreira, Luiz Pires, Guilherme Routar, and Sérgio Nunes. Merging datasets for aggressive text identification. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 128–139, 2018.
- Paula Fortuna, Juan Soler, and Leo Wanner. Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. In *Proceedings of the 12th language resources and evaluation conference*, pages 6786–6794, 2020.
- Paula Fortuna, Juan Soler-Company, and Leo Wanner. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524, 2021.
- Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Real-toxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3356–3369, 2020.

- Nicolas Gontier, Pau Rodriguez, Issam Laradji, David Vazquez, and Christopher Pal. Language decision transformers with exponential tilt for interactive text environments. *arXiv:2302.05507*, 2023.
- Reginald H Gonzales. Social media as a channel and its implications on cyber bullying. In *DLSU Research Congress*, pages 1–7, 2014.
- Shloak Gupta, S Bolden, Jay Kachhadia, A Korsunskaya, and J Stromer-Galley. Polibert: Classifying political social media messages with bert. In *Social, cultural and behavioral modeling (SBP-BRIMS 2020) conference. Washington, DC*, 2020.
- Georgios Hadjiharalambous, Kacper Beisert, and Joemon M Jose. End-to-end hierarchical approach for emotion detection in short texts. In *Responsible Data Science: Select Proceedings of ICDSE 2021*, pages 1–12. Springer, 2022.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34:1273–1286, 2021.
- Shagun Jhaver, Larry Chan, and Amy Bruckman. The view from the other side: The border between controversial speech and harassment on kotaku in action. *First Monday*, 23(2):4, 2018.
- Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling BERT for natural language understanding. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4163–4174, 2020.
- Google Jigsaw. URL <https://www.perspectiveapi.com/#/home>.
- Dacher Keltner, Lisa Capps, Ann M Kring, Randall C Young, and Erin A Heerey. Just teasing: a conceptual analysis and empirical review. *Psychological bulletin*, 127(2):229, 2001.
- Muhammad US Khan, Assad Abbas, Attiqah Rehman, and Raheel Nawaz. Hateclassify: A service framework for hate speech identification on social media. *IEEE Internet Computing*, 25(1):40–49, 2020.
- Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A large self-annotated corpus for sarcasm. In *proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.
- Haesoo Kim, HaeEun Kim, Juho Kim, and Jeong-woo Jang. When does it become harassment? An investigation of online criticism and calling out in Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–32, 2022.

- Taehyeon Kim, Jaehoon Oh, Nak Yil Kim, Sangwook Cho, and Se-Young Yun. Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 2628–2635. International Joint Conferences on Artificial Intelligence Organization, 2021.
- Robin M Kowalski. “I was only kidding!”: Victims’ and perpetrators’ perceptions of teasing. *Personality and Social Psychology Bulletin*, 26(2):231–241, 2000.
- Robin M Kowalski, Gary W Giumetti, Amber N Schroeder, and Micah R Lattanner. Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. *Psychological bulletin*, 140(4):1073, 2014.
- Robert V. Kozinets. The field behind the screen: Using netnography for marketing research in online communities. *Journal of Marketing Research*, 39(1):61–72, 2002. doi: 10.1509/jmkr.39.1.61.18935.
- Robert V Kozinets. *Netnography: redefined*. Sage, 2015.
- Mateusz Lango and Jerzy Stefanowski. What makes multi-class imbalanced problems difficult? an experimental study. *Expert Systems with Applications*, 199:116962, 2022.
- Loïc Lannelongue, Jason Grealey, and Michael Inouye. Green algorithms: quantifying the carbon footprint of computation. *Advanced science*, 8(12):2100707, 2021.
- Deborah Roth Ledley, Eric A Storch, Meredith E Coles, Richard G Heimberg, Jason Moser, and Erica A Bravata. The relationship between childhood teasing and later interpersonal functioning. *Journal of Psychopathology and Behavioral Assessment*, 28:33–40, 2006.
- Kuang-Huei Lee, Ofir Nachum, Mengjiao Sherry Yang, Lisa Lee, Daniel Freeman, Sergio Guadarrama, Ian Fischer, Winnie Xu, Eric Jang, Henryk Michalewski, et al. Multi-game decision transformers. *Advances in Neural Information Processing Systems*, 35:27921–27936, 2022.
- Jiacheng Li, Yujie Wang, and Julian McAuley. Time interval aware self-attention for sequential recommendation. In *Proceedings of the 13th international conference on web search and data mining*, pages 322–330, 2020.
- Courtney Subramanian Liv McMahon, Zoe Kleinman. Meta to replace “biased” fact-checkers with moderation by users, Jan 2025. URL <https://www.bbc.co.uk/news/articles/clly74mpy8klo>.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The FLAN collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR, 2023.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv:1711.05101*, 2017.
- Ariadna Matamoros-Fernández and Johan Farkas. Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, 22(2):205–224, 2021.
- Meta and Kaplan, Joel. More speech and fewer mistakes. <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>, January 2025. Newsroom Post.

- Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning-based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40, 2021.
- Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678, 2022.
- Raymond T Mutanga, Nalindren Naicker, and Oludayo O Olugbara. Hate speech detection in Twitter using transformer methods. *International Journal of Advanced Computer Science and Applications*, 11(9), 2020.
- Usman Naseem, Imran Razzak, and Ibrahim A Hameed. Deep context-aware embedding for abusive and hate speech detection on Twitter. *Aust. J. Intell. Inf. Process. Syst.*, 15(3):69–76, 2019.
- Esteban Ortiz-Ospina and Max Roser. The rise of social media. *Our world in data*, 2023.
- Keiron O’Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv:1511.08458*, 2015.
- Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523, 2021.
- Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. Leveraging intra-user and inter-user representation learning for automated hate speech detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 118–123, 2018.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Nils Reimers and Iryna Gurevych. all-mpnet-base-v2. <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>, 2021. Accessed: 2025-11-14.
- Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015. URL <https://arxiv.org/abs/1412.6550>.
- Dhruv Sahnan, Snehil Dahiya, Vasu Goel, Anil Bandhakavi, and Tanmoy Chakraborty. Better prevent than react: Deep stratified learning to predict hate intensity of Twitter reply chains. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 549–558. IEEE, 2021.
- Joni Salminen, Maximilian Hopf, Shammur A Chowdhury, Soon-gyo Jung, Hind Almerakhi, and Bernard J Jansen. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10:1–34, 2020.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv:1910.01108*, 2019.

- Vishwam Sankaran, Nov 2022. URL <https://www.independent.co.uk/independentpremium/world/elon-musk-twitter-layoffs-moderators-b2224818.html>.
- Shabnoor Siddiqui, Tajinder Singh, et al. Social media its impact with positive and negative aspects. *International journal of computer applications technology and research*, 5(2):71–75, 2016.
- Ge Song, Yunming Ye, Xiaolin Du, Xiaohui Huang, and Shifu Bie. Short text classification: a survey. *Journal of multimedia*, 9(5), 2014.
- Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. BERT4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1441–1450, 2019a.
- Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 4323–4332, 2019b.
- Stuart A. Thompson and Kate Conger. Meet the next fact-checker, debunker and moderator: You, Jan 2025. URL <https://www.nytimes.com/2025/01/07/technology/meta-facebook-content-moderation.html>.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023.
- Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models, 2019.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Bertie Vidgen, Tristan Thrush, Zeerak Talat, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 1667–1682, 2021b.
- Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do NLP models know numbers? Probing numeracy in embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5307–5315, 2019.
- Jie Wang, Alexandros Karatzoglou, Ioannis Arapakis, Xin Xin, Xuri Ge, and Joemon M. Jose. Sparks of surprise: Multi-objective recommendations with hierarchical decision transformers for diversity, novelty, and serendipity. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24*, page 2358–2368, 2024. Association for Computing Machinery. doi: 10.1145/3627673.3679533.

- Jie Wang, Alexandros Karatzoglou, Ioannis Arapakis, Joemon M. Jose, and Xuri Ge. Beyond accuracy: Decision transformers for reward-driven multi-objective recommendations. *IEEE Transactions on Knowledge and Data Engineering*, 37(9):5004–5016, 2025. doi: 10.1109/TKDE.2025.3582506.
- Qiong Wang, Ruilin Tu, Yihe Jiang, Wei Hu, and Xiao Luo. Teasing and internet harassment among adolescents: The mediating role of envy and the moderating role of the Zhong-Yong thinking style. *International Journal of Environmental Research and Public Health*, 19(9):5501, 2022.
- Tianyi Wang, Ke Lu, Kam Pui Chow, and Qing Zhu. Covid-19 sensing: negative sentiment analysis on social media in china via bert model. *Ieee Access*, 8:138162–138169, 2020a.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020b.
- Xia Wang, Chunling Yu, and Yujie Wei. Social media peer communication and impacts on purchase intentions: A consumer socialization framework. *Journal of interactive marketing*, 26(4):198–208, 2012.
- Oliver Warke, Joemon M Jose, Jan Breitsohl, and Jie Wang. Capturing the spectrum of social media conflict: A novel multi-objective classification model. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pages 215–225, 2024.
- Janis Wolak, Kimberly J Mitchell, and David Finkelhor. Does online harassment constitute bullying? an exploration of online harassment by known peers and online-only contacts. *Journal of adolescent health*, 41(6):S51–S58, 2007.
- Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Wikipedia Talk Labels: Personal Attacks. 2 2017a. doi: 10.6084/m9.figshare.4054689.v6. URL [https://figshare.com/articles/dataset/Wikipedia\\_Talk\\_Labels\\_Personal\\_Attacks/4054689](https://figshare.com/articles/dataset/Wikipedia_Talk_Labels_Personal_Attacks/4054689).
- Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399, 2017b.
- Lan Xia. Effects of companies’ responses to consumer criticism in social media. *International Journal of Electronic Commerce*, 17(4):73–100, 2013.
- Chuanpeng Yang, Yao Zhu, Wang Lu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, and Yiqiang Chen. Survey on knowledge distillation for large language models: methods, evaluation, and application. *ACM Transactions on Intelligent Systems and Technology*, 2024.
- Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv:1612.03928*, 2016.
- Mark Zuckerberg. More speech and fewer mistakes. Facebook Video, January 2025. URL <https://www.facebook.com/watch/?v=1525382954801931>.