CDB: A Unified Framework for Hope Speech Detection Through Counterfactual, Desire and Belief

Tulio Ferreira Leite da Silva[1,2], Gonzalo Freijedo Aduna[3], Farah Benamara[4,5], Alda Mari[3], Zongmin Li[2,6,7], Li Yue[6], Jian Su[6]

[1] University of Sao Paulo, Brazil
[2] CNRS@CREATE LTD, Singapore
[3] Institut Jean Nicod CNRS/ENS/EHESS/PSL University
[4] IRIT, Université de Toulouse, CNRS, Toulouse INP, Toulouse, France
[5] IPAL, CNRS-NUS-A*STAR, Singapore
[6] Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
[7] School of Computer Science and Engineering, Nanyang Technological University, Singapore

Correspondence: farah.benamara@irit.fr

Abstract

Computational modeling of user-generated desires on social media can significantly aid decision-makers across various fields. Initially explored through wish speech, this task has evolved into a nuanced examination of hope speech. To enhance understanding and detection, we propose a novel scheme rooted in formal semantics approaches to modality, capturing both future-oriented hopes through desires and beliefs and the counterfactuality of past unfulfilled wishes and regrets.We manually re-annotated existing hope speech datasets and built a new one which constitutes a new benchmark in the field. We also explore the capabilities of LLMs in automatically detecting hope speech. To the best of our knowledge, this is the first attempt towards a language-driven decomposition of the notional category hope and its automatic detection in a unified setting.

1 Introduction

Hope is a notional category that lies at the crossroads of desire, belief and intention as the psychological (Snyder, 2000; Eliott and Olver, 2002), philosophical (Bloeser and Stahl, 2022; Godfrey, 2012) and the linguistic literature (Heim, 1992; Ruffman et al., 2003; Portner and Partee, 2008; Anand and Hacquard, 2013; Grano, 2017) have independently highlighted.

As an expression of will and intention, hope functions as motivation for action (Bratman, 1987). It forms a mental attitude that contributes to the rational behavior of an agent to foresee and plan new strategies to fulfill what people would like to see happen or to have happened. Examining hopes offer valuable insights for decision-makers in various domains ranging from customer service and marketing (Carlos and Yalamanchi, 2012), politics (Hirschman, 1970), and decision-making processes (Hammond et al., 1998; Guo, 2023).

While acquiring, modeling, and reasoning with desires and beliefs have been extensively studied in artificial intelligence (Cohen and Levesque, 1990; Georgeff et al., 1999), automatic extraction of hope linguistically expressed in texts has received comparatively less attention in the NLP literature. The first attempt in this direction was the seminal work by Goldberg et al. (2009), followed by Ramanand et al. (2010) and Chang et al. (2013), on binary wish classification where a wish is defined as "a desire or hope for something to happen". Recently, the novel task known as *hope speech detection* (hereafter H*p*SD) has gained significant interest in the NLP community where many shared tasks have been organized, like the LT-EDI series @ACL and RANLP (Chakravarthi et al., 2021, 2022b, 2023). Balouchzahi et al. (2023b) and Balouchzahi et al. (2023a) go beyond binary classification and proposed a multiclass model for a nuanced definition of hope and regret, respectively.

Hope is both past (cf. (1)) and future (cf. (2)) oriented, and reveals a belief-based expectation or a desire-based projection that can be either realized, or that can no longer become actual.

(1) US surveillance drones shouldn't have been operating where they were told not to operate.

(2) I am very interested into drones and I'm currently looking into the prospect of buying a drone in Singapore.

Each existing hope dataset has its own annotation scheme, characterizing hope either as: (a) *only past-oriented* to address regrets or (b) *only future-oriented* to account for general or specific

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 4448–4463

hopefulness, optimism as well as wishes. In addition, existing studies focus on *the positive aspect of hope* where hope is viewed either as a counterpoint to hate (i.e., social awareness, minority protection, etc.) or as a way to offer support and inspiration (Chakravarthi, 2020). Balouchzahi et al. (2023b)'s taxonomy aimed to go beyond positivity by introducing unrealistic expectations. However, annotators often consider them as regret therefore labeling them as not-hope since they are not future-oriented (e.g., "*Wish there was a ban on fireworks*").

This paper addresses these drawbacks by disentangling hope orientation (past and future) from the valence of the content of the hope (we consider hope as both directed towards positive and negative goals). We rely on linguistic clues at the semantic / pragmatic interface to identify the relevant categories and propose a broader and language-driven design of the landscape of hope that covers the scope of previous works, while gaining in informativity. Our contributions are as follows:

1. **A novel language-driven decomposition of the notional category HOPE** that articulates its building blocks by distinguishing between **C**ounterfactual (as past oriented hope), and future oriented hope on the one hand, and between **D**esire-based and **B**elief-based hope in the realm of future orientation (Condoravdi, 2002; Portner, 2009a; Anand and Hacquard, 2013; Grano, 2017). We hereafter refer to our model Counterfactual-Desire-Belief as the CDB model (cf. Section 3).

2. **An English dataset of about 4,370 texts annotated according to this characterization**. We manually re-annotated a subset of existing hope datasets and built a new one while merging them under the same scheme rooted in formal semantics approaches to modality (cf. Section 4). This forms a new benchmark in the field.[1]

3. **A set of experiments to detect hope using both transformers and large language models** relying on various prompting strategies (cf. Section 5). LLMs capabilities have been explored in sentiment analysis, emotion and offensive language detection (Zhang et al., 2024; Li et al., 2023). The use of open-source LLMs for H*p*SD is new.

2 Hope Speech in NLP

2.1 Hope Speech Datasets

Table 1 summarizes existing H*p*SD datasets.[2] Most of them are in English with a focus on social media generated content. They can be grouped into two main categories. The first one views hope as opposite to hate with the aim of moderation through the detection of positive content. Palakodety et al. (2020) rely on n-grams and automatically build a corpus tagged as pro-war and pro-peace intent in nuclear conflict escalation. HopeEDI (Hope Speech Detection Dataset for Equality, Diversity, and Inclusion) (Chakravarthi, 2020) is composed of Youtube comments and casts the problem into a binary task (hope vs. not-hope) where hope promotes positivity within minorities, while not-hope encompasses hate speech and everything else. A subset of this dataset has been manually re-annotated by Aggarwal et al. (2023) to account for neutral cases (non English texts, lack of context).

The second group views hope beyond hate and proposes the first attempts to implement a theoretical framework leveraging insights from psychology. PolyHope contains tweets about women's child abortion, black people's rights, religion, and politics. Its annotation scheme considers four categories (Balouchzahi et al., 2023b): (i) not-hope, (ii) generalized hope, (iii) unrealistic hope, and (iv) realistic hope. ReDDit (Regret Detection and Domain Identification from Text) (Balouchzahi et al., 2023a) on the other hand distinguishes between regrets by action vs. inaction in posts from "regret", "regretfulparents", and "confession" subreddits. Finally, the WISH corpus (Goldberg et al., 2009) offers a set of 7,6K sentences from consumer product reviews and political discussion board postings manually annotated for wish vs. not-wish following a set of lexico-syntactic wish patterns (e.g., `would be better if X`). Table 2 provides a summary of annotation categories and their definitions in existing hope datasets, as given by the referenced papers.

Our work is in line with the PolyHope and ReDDit view that a variety of dimensions have to be captured, while covering the same large scope of

[1] The annotation guideline and a subset of the dataset are available as a supplementary material. The dataset will be made available to the research community.

[2] Studies that focus on the role of desire and belief in modelling emotion and sentiment (e.g., (Jia et al., 2022; Xu et al., 2024)) are out of the scope of this paper.

Corpus	Instances	Language	Hate-speech related	Manually Annotated
WISH Corpus (Goldberg et al., 2009)	7,614	English	-	Yes
Pro-War/Pro-Peace Dataset* (Palakodety et al., 2020)	2,047,851	English, Hindi	Yes	-
HopeEDI (Chakravarthi, 2020)	59,354	English, Tamil, Malayalam	Yes	Yes
HopeEDI with Neutral (Aggarwal et al., 2023)	23,003	English	Yes	Yes
PolyHope (Balouchzahi et al., 2023b)	8,256	English	-	Yes
ReDDit (Balouchzahi et al., 2023a)	3,425	English	-	Yes

Table 1: State of the art datasets for hope speech detection. "*": Not publicly available.

Corpus	Categories	Definition
WISH	Wish	"A desire or hope for something to happen." + a set of wish templates
PolyHope	Not-Hope	"Not indicate hope, wish, desire, or future-oriented expectation."
	Generalized Hope	"Expresses a general hopefulness and optimism not directed towards any specific event or outcome."
	Realist Hope	"Expecting something reasonable, meaningful, and possible thing to happen (there is every possibility that this will happen)."
	Unrealistic Hope	"Wishing for something to become true, even though the possibility of happening is remote, significantly less, or even zero."
ReDDit	Not-Regret	"Does not convey any type of regret."
	Regret by Action	"Regretting a decision or choice or an action one has done in the past."
	Regret by Inaction	"A regret caused by a lack of decision or failure in doing something."

Table 2: Annotation categories and their definitions in existing hope datasets, following Goldberg et al. (2009) (WISH), Balouchzahi et al. (2023b) (PolyHope) and Balouchzahi et al. (2023a) (ReDDit).

HopeEDI. Our proposal distinguishes itself in that it provides a more fine-grained classification and a unified view that offers several advantages: (1) From a more theoretical point of view, it restates regret in more general terms as counterfactual hope (as hope held in the past and no longer actualizable), and articulates hope around linguistic dimensions manifested across well-studied categories and the modal / temporal combinations. (2) It will provide to the research community a corpus for H*pSD* that encompasses existing manually annotated corpora. (3) It would allow us to compile a list of markers and other signals, which would also benefit automatic detection. Indeed, adopting a linguistic-based analysis allows us to rely on well-established classes of linguistic clues (phrasal mood, attitudes and modals and a combination of those in with and without tense) that guide our classification in a reliable and easily reproducible manner.

2.2 Automatic Detection

First approaches relied on lexico-syntactic patterns injected as features in supervised learning models (Goldberg et al., 2009; Ramanand et al., 2010; Dong et al., 2013; Chang et al., 2013). Deep learning such as LSTM and BiLSTM (Chakravarthi et al., 2022a) and transformers specifically have shown to be particularity useful in binary classification. For example, Zhu (2023) used DistillBert and obtained the best score at HOPE@IberLEF 2023 in English (Ureña López et al., 2023). Sidorov et al. (2023) experiment with multiclass classification of hopes (PolyHope) and regrets (ReDDit) datasets relying on GloVe embeddings. Finally, the use of LLMs (ChatGPT) has been reported in Ngo and Tran (2023) for H*p*SD in Spanish.

In this paper, we experiment with transformers but also and for the first time various open-source LLMs with different prompting strategies. Our models have been evaluated on a unified hope speech dataset that spans over multiple domains which allow for measuring models generalization across datasets in detecting fine-grained manifestations of hope. Several efforts have been carried out for cross-dataset experiements in many NLP tasks such as hate speech (Fortuna et al., 2021) and fallacy detection (Helwe et al., 2024). Our work is the first attempt towards H*p*SD in a unified setting.

3 A Language-Driven Characterization of Hope

3.1 The CDB Model

Hope as a modal category. In our study, we envision the notion of hope as a *topic* of the message that is manifested in the text across a variety of *modal* expressions and the inferences that these expressions enhance.

Strictly speaking, the lexical term *hope* belongs to the grammatical class of propositional attitudes

(e.g. (Hamblin, 1973; Portner, 2018; Condoravdi and Lauer, 2012; Anand and Hacquard, 2013; Grano, 2017)) and in particular to the bouletic class (*want, wish, hope, like*, ...). However, at the message level, hope has a more complex range of meanings and manifestations, and can be viewed *as an expectation imbued with desire* (and in this case the belief is foregrounded) or as *a desire of an outcome that the speaker believes possible* (and, in this case, the desire is foregrounded) (see Anand and Hacquard (2013)). This distinction between two types of hope (belief and desire based) is enhanced by specific *modal expressions* (e.g. *must, might, should, sorry, believe, want*, ...). Modal expressions affect the truthiness of the sentence, and reveal stances and postures of the speakers.

Our annotation of hope as a topic is based on the formal semantic approach by Portner (2009b) that distinguishes a variety of modal meanings that we build on (see Figure 1): **Epistemic** modals are based on belief/knowledge (e.g. *John **must** be at home, as far as I know*), **Deontic** on rules (e.g. *You **must** vote, according to the rules*), **Volitional** modals (e.g. *want*) are based on desires while **Teleological** on plans and goals (e.g. *If you want to go to Harlem, you **must** take the A train*).

EPISTEMIC	PRIORITY		TELEOLOGICAL
Belief based	Deontic Rule based	Volitional Desire based	Goal & Plan based

Figure 1: Modal meanings.

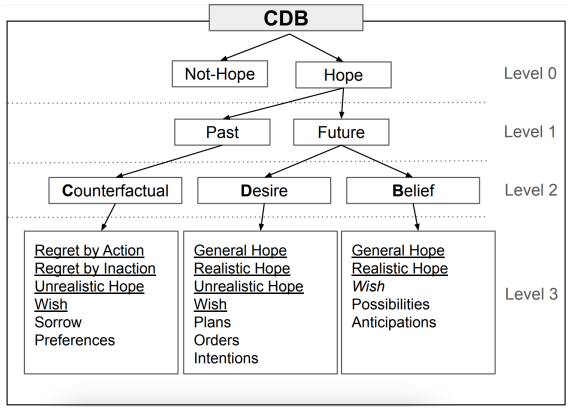

Figure 2: The Counterfactual-Desire-Belief (CDB) model. The wish under belief is in italic font to signal that we newly consider belief-grounded wishes.

Hope annotation and modal meaning mapping. We map these modal meanings into hope categories. This avoids arbitrariness: Modal meanings have precise manifestations, and by controlling these clues, we avoid subjective annotation of the topic hope at the message level. In particular, the Epistemic and Deontic modal categories that relate to facts and legal (or legal-like) rules map into the hope-as-belief (hereafter BELIEF). On the other hand, Volitional and Teleological that rely on desires and inherently involve preferences map into the hope-as-desire category (hereafter DESIRE).

As the overall topic of the messages are always enhanced by precise lexical information and to avoid projecting the annotator subjectivity, we have established **clear annotation rules based on grammatical and lexical manifestations of modality**. Modality surfaces across the lexical, phrasal and grammatical domains. In particular, in our study, we have considered the following categories. (i) Sentential mood, and most specifically imperatives and their variety of interpretations (Condoravdi and Lauer, 2012) (orders, wishes, invitations, advices...) and optatives (*please, if ... only*) (Grosz, 2012; Portner, 2018), (ii) modals auxiliaries and verbs (*could, should*), adverbs (*hopefully, lukily, ...*), within specific temporal setting (past and future tenses) and within larger phrases (*I wish I could, should have been, had I known that*), (Portner, 2009a)), and (iii) attitudes (*wish*, *anticipate*, *hope*), also across different temporal settings (Anand and Hacquard, 2013; Villalta, 2008; Grano, 2017). Once the modal expressions are identified along with their lexical meanings, the hope category is easily assigned.

The temporal dimension. Finally, as Figure 2 shows, we distinguish between future and past orientation (which we label COUNTERFACTUAL). This past-oriented category encompasses both beliefs and desires that were held in the past and are no longer realizable. While future orientation and expressions of hopes that are still realizable are more useful to decision makers, considering past hopes can still give an indication of the stances and postures of the users. We have nonetheless decided to conflate desires and beliefs in the counterfactual category to avoid unnecessary complexity.

Our linguistic characterization of hope is complementary to (Saurí and Pustejovsky, 2009) in that it explicitly focuses on desire and expectations and not on factuality (and lack thereof) as a general category cutting across different notional domains beyond hope. Our approach extends the Belief-Desire-Intention model (Georgeff et al., 1999) by introducing an explicit temporal dimensions, and

recasting the annotation with a language-driven approach. While these are well-established models that *prima facie* might appear related to ours, our modal deepens the investigation of belief and volition domains beyond previous achievements.

3.2 The Proposed Annotation Scheme

CDB embeds in a uniform way all the publicly existing datasets by spelling out precisely the concepts covered and extending beyond categories usually treated in the literature (underlined in Figure 2) proposing a clear-cut distinction between belief-based and desires based-concepts as grounding different aspects of hope, as shown in Table 3. Our categories are:

(a) Past-oriented hope

COUNTERFACTUAL. It refers to a desire, wish, or longing for reality or outcome different from what is the case. Counterfactual hope can describe hopes that could have been realistically become true as well as hope that, even in the past, could have never been realized. This can emerge as regret (4), sorrow (3), or beliefs that alternative scenarios should be the case. Its hallmark is past orientation with attitude verbs or modal auxiliary.

(3) *So bummed I am out of town. Was hoping to catch this show. Enjoy!*

(4) *if i knew then what i know now i would have paid the extra to get an ipod avoided the zen*

(b) Future-oriented hope

DESIRE. It refers to the expression of a mere wish, as an emotion (5); plans, which are desires to act next (6) ; and imperatives, which issue orders or desires for others to act (7).

(5) *overall it is a great unit hopefully creative labs or some other vendor will soon come out with some accessories for it.*

(6) *i ve been looking to buy a digital camera for a long time and v finally decided that now was the time.*

(7) *right please go and check out a book from any library.*

BELIEF. This category encompasses expressions of an epistemic state. Even in scientific contexts, rational hope inherently involves an expectation that events will unfold as epistemically anticipated (Heim, 1992). Desire is present but it is backgrounded as illustrated in (8). Verbs such as *should*, *would*, and *could* often carry an implicit epistemic presupposition (Portner, 2009a; Rubinstein et al., 2013) as in (9). The desire component of belief-based hope can be enforced into deontological considerations, as in (10). Likewise, belief can also encompass unrealistic hopes (cf. (11) that does express a desire).

(8) *There is a power outage on the west side of town around Spanish Fork High School [...] We anticipate the power will be restored in 2-3 hours.*

(9) *so who hear thinks i should keep the money*

(10) *there should be no abortions or pulling the plug on a dead woman from Florida*

(11) *Maybe if you're very good a crab will one day wear your skull too.*

(c) NOT-HOPE

It encompasses everything else, including narrations (12) among others like compliments.

(12) *officers discovered that the man was standing beside them in the police line shouting please come out and give yourself up*

4 Data and Annotation

4.1 Data Sources

– Re-annotating existing datasets. We relied on three publicly available hope datasets to construct our unified benchmark:[3] **WISH** (Goldberg et al., 2009) about politics and products, **PolyHope** (Balouchzahi et al., 2023b) and **HopeEDI with neutral** (Aggarwal et al., 2023) about minority protection. We randomly selected 3,092 texts from these corpora while balancing instances from each original class per dataset/topic. Original annotations have been removed to avoid bias.

– A new dataset. To ensure diversity in terms of topics, we built **HopeDrone**, a dataset about drone-related discussions in social media. Monitoring discourse about drones has increasingly attracted the attention of governments seeking to develop more effective public policies (Hall, 2015; Hwang et al., 2019). Therefore, understand people's hopes and expectations can also contribute to broader decision-making applications of H*p*SD. We

[3]The ReDDit dataset is available upon request but after several requests, we could not have access to the data.

Corpus	Categories	Temporal Dimension		Modal Dimension		
		Future	Past	Sentence Mood	Modal Verbs, Particles, Adverbs	Attitude
Wish Corpus	Wish	✓	-	✓	✓	✓
PolyHope	General Hope	✓	-	-	-	-
	Realist Hope	✓	-	-	-	-
	Unrealistic Hope	✓	-	-	-	-
ReDDit	Regret by Action	-	✓	-	-	-
	Regret by Inaction	-	✓	-	-	-
CDB (ours)	Counterfactual	-	✓	-	✓	✓
	Desire	✓	-	✓	-	✓
	Belief	✓	-	-	✓	✓

Table 3: The CDB model (ours) vs. existing models.

focused on surveillance, delivery, taxi services, and regulatory frameworks. These topics were identified through an initial analysis of social media discussions as particularly vibrant and directly impactful on policy-making considerations.

We used a set of dedicated keywords (see Appendix A) to scrap data in English from both Twitter and Reddit which allowed us to stream 12,189 tweets and 12,001 posts, after removing retweets and duplicates. We randomly selected a subset composed of 1,216 tweets and 62 Reddit posts for the annotation campaign covering all the four topics about drones.

Overall, our corpus consists of 4,370 texts that have been manually (re-)annotated. Our data is diverse in terms of source, domain and text length ranging from 10 to 988 characters (cf. Table 4).

4.2 Annotation Procedure

Annotation consists in assigning to each instance one of the following four categories: COUNTERFACTUAL, DESIRE, BELIEF and NOT-HOPE. A class named NO DECISION has been used in case of indecision, e.g. lack of context, difficulties related to complex construction, etc.

Two annotators (master's degree and PhD students in linguistics) participated in the process. We performed a four-step annotation where an intermediate analysis of agreement and disagreement between the annotators was carried out: (1) Train annotators on 200 instances while sharing understanding of the theoretical framework and classification criteria, (2) Doubly annotate a second pool of 200 instances (Kappa K=72.70 (binary) and 60.49 (multiclass)) then discuss disagreement cases which helped in updating the manual, (3) Doubly annotate a third pool of 200 instances (K=77.00 (binary) and 70.85 (multiclass)). (4) Doubly annotate 2,000 instances (those used in steps 1-3 have been discarded). This results in a K= 67.89 for HOPE/NOT-HOPE/NO DECISION and 65.86 for the 5 classes annotations. After removing the NO DECISION (694 messages), the Kappa increases to 74.88 for binary classification vs. 70.46[4] for 4 classes. The Kappa being good given the subjectivity of the task, an additional 1,770 instances have been annotated by consensus. All the annotations were done manually, without relying on LLMs.

4.3 Results

Qualitative Results. As shown in the confusion matrix in Appendix B.1, common disagreements concern COUNTERFACTUAL vs. DESIRE (28.71%), BELIEF vs. NOT-HOPE (23.05%) and DESIRE vs. NOT-HOPE (12.71%). Main reasons of these disagreement are cases related to complex constructions where hope was being reported (e.g., the speaker is thanking the addressee for their desire) and reinterpretations of the modal meanings in specific contexts (see Appendix B.2 for details). Another issue is related to cases where more than one category was being expressed. To disambiguate these cases, annotators were asked to assign, in addition to the main hope class, a secondary one (about 14.4% instances). This allows to determine if there was any hierarchy between multiple hopes expressed within a sentence. For instance, if a sentence starts with a desire followed by two beliefs, the correct classification would be belief. However, the initial desire might heavily

[4]This kappa is lower compared to PolyHope (0.82) and ReDDIT (0.78) for two reasons: (a) Our data has a more diverse topics and genres which trigger complex linguistic phenomena, (b) Our annotation is more fine grained covering past and future-oriented hope.

Category	HopeDrone	PolyHope	WISH Corpus	HopeEDI	Total
Counterfactual	15 (0.34%)	112 (2.56%)	62 (1.42%)	14 (0.32%)	**203 (4.65%)**
Desire	224 (5.13%)	682 (15.60%)	360 (8.24%)	113 (2.59%)	**1,379 (31.56%)**
Belief	390 (8.92%)	202 (4.62%)	226 (5.17%)	103 (2.36%)	**921 (21.08%)**
Not-Hope	649 (14.85%)	279 (6.39%)	569 (13.02%)	370 (8.47%)	**1,867 (42.72%)**
Total	**1,278 (29.24%)**	**1,275 (29.18%)**	**1,217 (27.85%)**	**600 (13.73%)**	**4,370 (100%)**

Table 4: The CDB dataset.

influence the classification. Therefore, having both primary and secondary classifications helped us understand these nuances and agree on the category that was most prominent.

Quantitative Results. The final CDB annotations have been assigned by consensus, solving all the NO DECISION. Among them 31.56% are DESIRE, 21.08% BELIEF while COUNTERFACTUAL is the minority class (cf. Table 4).

We further analyze the mapping between our new CDB annotations and the original ones in the original. First, we compare the CDB re-annotations of the WISH corpus in the multiclass (cf. confusion matrix in Figure 3) setting. The matrix corresponds to 978 instances that have been doubly annotated, all disagreements have been solved by consensus. We then compare the CDB re-annotations of the PolyHope (cf. matrix in Figure 4 that corresponds to 1,022 instances that have been doubly annotated, all disagreements have been solved by consensus).

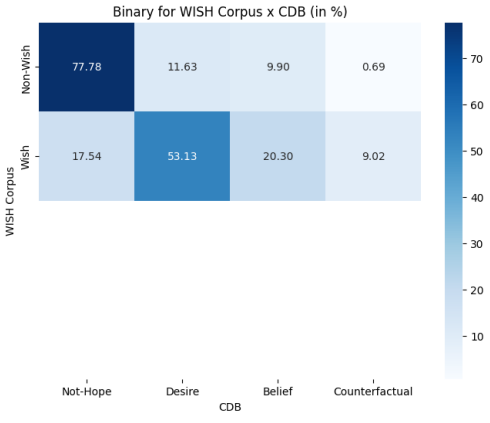

Figure 3: Comparison between CDB and WISH multiclass annotations.

When looking into binary classification HOPE vs. NOT-HOPE where desires, beliefs and counterfactuals are merged under the same top level HOPE class, around 83% of HOPE instances were annotated with a similar category as the original one.

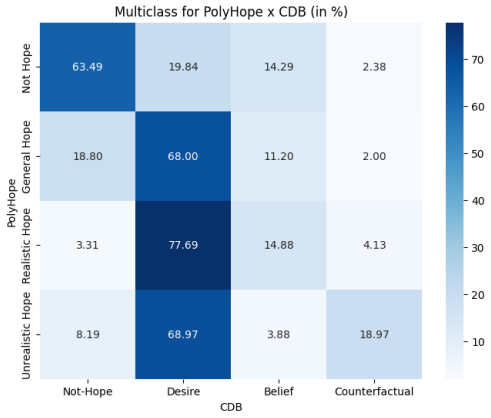

Figure 4: Comparison between CDB and PolyHope multiclass annotations.

This percentage decreases to 63.49% and 77.78% for NOT-HOPE. However, when looking into fine-grained annotations of the WISH corpus, it is interesting to observe that most of the wish class corresponds to desires (53.13%) but also to beliefs (20.30%) and to a little extent to counterfactuals (9.02%). On the other hand, 11.63% of not wishes have been re-annotated as desires. Similarly, in the PolyHope subset, desires correspond to 77.69% of realistic and 68.97% unrealistic hope. More importantly, only 18.97% of unrealistic hopes have been re-annotated as counterfactuals. This shows that a unified view is attainable and timely (see Figure 2).

5 Hope Speech Detection

5.1 Experimental Settings

We randomly split the corpus into train and test sets (85%-15% test sets) while balancing the distribution of CDB categories as well as the source of data among our four datasets. The data is imbalanced with 43.03% (resp. 40.98%) of NOT-HOPE in the train (resp. test). More importantly, the train is composed of instances from each dataset with a similar distribution: 30.19%, 27.45%, 26.21% and 16.15% are from HopeDrone, PolyHope, WISH and HopeEDI respectively, the last one being only

present in the train. Appendix C.1 details the distribution of the hope categories and the source of the data in the train/test sets. We experiment with the following models (see Appendix C.2 and C.3 for their description and the used parameters):

- **Transformers.** We fine-tune $\text{BERT}_{\text{base}}$ (Devlin et al., 2019) and $\text{RoBERTa}_{\text{base}}$ (Liu et al., 2019) in binary and multiclass classification settings.

- **Non fine-tuned LLMs.**[5] We rely on 6 open source LLMs with different sizes from the HuggingFace library without specific training: Llama3_{8B} and Llama3_{70B} (AI@Meta, 2024), $\text{Mistral}_{\text{7B - Instruct - v0.3}}$ (Jiang et al., 2023), Aquila2_{7B}[6], Starling-LM $_{\text{7B-beta}}$(Zhu et al., 2023), and $\text{Vicuna}_{\text{13B-v1.5}}$ (Zheng et al., 2023). As LLMs may have different output formats, we consider a class correct if it exactly corresponds to of one of our categories. We employ the following baseline prompting strategies to generate the corresponding CDB labels[7] (see Appendix C.4): **Zero-shot** that presents the model with the task, the definition of each of and the sentence to be labeled, and **Few-shot in context learning (ICL)** where we augment the zero-shot instructions with two examples per category randomly sampled from the train set, for a total of 8 examples.

- **Fine-tuned LLMs.** An analysis of outside of categories rate (OC) (i.e., the number of predictions that do not correspond to the classes the LLM is instructed to generate out of the total number of instances in the test set) shows that Mistral was the most stable model with an OC rate of 0.003% (resp. 0%) in the zero-shot (resp. few-shot) setting. We therefore fine-tune Mistral with LoRA (Hu et al., 2022).

To avoid bias, all LLMs share the same prompts and parameters. Both transformers and LLMs have been tested on the same test set and run 3 times with different seed of examples in the few-shot settings. In the following, we report the averaged precision, recall, and macro-F1 scores together with standard deviation.

[5]Due to institutional constraints, we only experimented with open source LLMs.

[6]https://github.com/FlagAI-Open/Aquila2

[7]Following Helwe et al. (2024), binary prediction results have been obtained by extrapolating multiclass predictions.

	Binary	Multiclass
Supervised Methods		
BERT	86.46 (0.00)	**77.06** (0.00)
RoBERTa	**86.67** (0.00)	74.47 (0.00)
Zero-shot Setting		
Aquila2	38.32 (0.00)	22.27 (0.05)
Llama3_{8B}	47.95 (0.20)	32.15 (0.91)
Llama3_{70B}	**72.96** (0.00)	**55.25** (0.00)
Mistral	68.88 (0.34)	54.49 (0.86)
Starling	41.81 (1.76)	37.51 (1.09)
Vicuna	44.22 (0.40)	32.51 (0.76)
Few-shot in Context Setting		
Aquila2	48.57 (6.61)	31.35 (1.30)
Llama3_{8B}	56.14 (7.38)	43.87 (4.60)
Llama3_{70B}	**73.02** (3.14)	48.35 (8.78)
Mistral	57.79 (2.67)	48.30 (3.85)
Starling	69.49 (4.37)	**56.51** (4.12)
Vicuna	58.92 (1.87)	44.51 (0.46)
Fine-tuned LLM		
Mistral	**88.80** (0.31)	**82.81** (0.27)

Table 5: Macro F1-score results. Best scores per setting are in bold. Standard deviations are between brackets.

5.2 Results

Table 5 shows our results. Transformer models outperform non fine-tuned LLMs, the best model being BERT with a macro F1-score of 77.06 in multiclass and RoBerta with 86.67 in binary classification. Although the dataset is imbalanced, it is interesting to note that all hope categories achieved similar results with a F1-score of 77.97, 70.39, 74.67, 85.23 for DESIRE, BELIEF, COUNTERFACTUAL and NOT-HOPE respectively. Regarding LLMs, best small sized models are Starling with 56.51 in the ICL setting and Mistral with 68.88 in the zero-shot scenario. Two main conclusions can be derived: (1) Although the larger Llama is the best, smaller models achieved comparable performances (see Mistral zero-shot). The results are even worse in ICL where Starling beats Llama3_{70B} in multiclass classification, (2) Fine-tuning Mistral was very productive achieving comparable results with supervised transformers.

Finally, we also analyzed the results of our best models per dataset and per class in the multiclass configuration. The scores are stable across datasets which shows that the model is able to deal with hope manifestations in different corpus genres and domains. However, when analyzing these results, it is important to note that our CDB model introduces a new benchmark that has not been previously explored, which makes direct comparisons with existing benchmarks inaccurate (cf. mappings in Table 3). In particular, the annotation campaign shows (cf. Figures 3 and 4) complete statistical

Examples	BERT	Mistral FT	Gold
(a) i wish i were an idol so people would make fun 24-second edits of all the times i've ever looked even a LITTLE sassy in public and set it to flo milli	Desire	Counterfactual	Counterfactual
(b) My problem is i expect a man to be obsessed with me from day 1 .. and if they not then i ghost them	Desire	Belief	Desire
(c) maybe you should look it up in the dictionary	Desire	Desire	Not-Hope
(d) "Thomas Scholars" could become a badge of honor - something kids could aspire to. Let Thomas pick his own Board of Trustees and personally interview the finalists.	Desire	Belief	Belief

Table 6: Error analysis of our best transformer and LLM models. Each predicted label is obtained by a majority vote over 3 runs.

	Mistral			Fine-Tuned Mistral		
	P	R	F1	P	R	F1
PolyHope						
Counterfactual	0.42	0.79	0.55	0.80	0.97	**0.88**
Desire	0.76	0.76	**0.76**	0.81	0.87	0.84
Belief	0.40	0.57	0.46	0.69	0.62	0.65
Not-Hope	0.67	0.19	0.30	0.80	0.64	0.71
WISH product						
Counterfactual	0.41	0.86	0.55	1.00	0.86	**0.92**
Desire	0.36	0.70	0.47	0.73	0.77	0.75
Belief	0.47	0.42	0.43	0.77	0.69	0.73
Not-Hope	0.84	0.43	**0.57**	0.89	0.90	0.89
WISH politics						
Counterfactual	0.33	0.78	0.46	0.94	0.89	**0.91**
Desire	0.61	0.82	**0.70**	0.89	0.79	0.84
Belief	0.57	0.49	0.51	0.79	0.81	0.80
Not-Hope	0.77	0.43	0.54	0.82	0.90	0.86
HopeDrone						
Counterfactual	0.00	0.00	0.00	1.00	1.00	**1.00**
Desire	0.36	0.79	0.48	0.70	0.68	0.69
Belief	0.57	0.45	0.49	0.80	0.75	0.77
Not-Hope	0.92	0.72	**0.81**	0.92	0.95	0.94

Table 7: Multiclass results (average of 3 runs) for Mistral and Fine-Tuned Mistral per class and dataset in terms of precision (P), recall (R), and macro F1-score (F1). Best F1 score per dataset is in bold font.

details about this mapping. For instance, DESIRE encapsulates all PolyHope categories.

5.3 Error Analysis

A manual error analysis of BERT and Mistral shows four main causes of misclassification (cf. Table 6): (1) Counterfactuals instances (example (a) in Table 6) classified as desire mainly because models seemed to give too much weight to the occurrence of *wish*, disregarding the counterfactual constructions. (2) The models, especially Mistral, are very sensitive to keywords in some of the labeling. For example, they often confuse desire with belief when the verb *to expect* appears (as in example (b)). Similarly, labeling desire by the appearance of the verb *wish*, disregarding the modal verb *should*, which is what introduces belief. (3) The models seem to have difficulties labeling advice and/or imperatives, usually classifying them as desire or belief, as in example (c). They also have problems with questions or rhetorical questions, especially if they include attitudinal verbs such as *wish* or *yearn*. Finally, (4) when several categories are expressed (like in (d)), the models seem to give preponderance to the first occurrence, regardless of whether it is the prevailing category according to the ground truth.

6 Conclusion

We proposed Counterfactual-Desire-Belief, a language-driven approach in the characterization of hope in texts that cover both modal (sentence mood, modal verbs, particles, adverbs and attitude) and temporal (past vs. future orientation) dimensions. Our framework offers a unified view of hope that allows to merge all existing hope speech datasets in a uniform way. We conducted an annotation campaign re-annotating three state of the art corpora as well as a new dataset, forming a new benchmark in the field. We also experimented with hope speech detection using both transformers and open source LLMs relying on standard prompting and new prompts that guide the LLM through score and topic instructions. We plan to extend the CDB model in a multimodal and multilingual settings.

Acknowledgment

This work has been supported by DesCartes: the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program. Alda Mari gratefully thanks ANR-17-EURE-0017 FrontCog. Tulio Ferreira Leite da Silva is thankful to Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) 23/1010-1 for their support.

Ethics Statement

The data used for conducting the experiments are composed of texts taken from three datasets publicly available to the research community. The new dataset about drone does not contain any offensive or abusive language.

Regarding the annotation campaigns, the annotators are co-authors of this paper and have been granted as part of their doctoral studies in linguistics. Both have a very strong experience in annotating subjective expressions in previous annotation campaigns (sentiment, emotion, irony, etc.).

Limitations

We utilized various open-source large language models in our experiments. It is important to acknowledge that these LLMs can exhibit biases and may encounter issues concerning token limit. Therefore, a critical approach should be adopted when interpreting the experimental outcomes.

All our data are in English. Hope speech being a subjective phenomena, language as well as culture may impact on the way hope can be expressed. Investigating these issues is left for future work and would provide a more comprehensive understanding of hope at the theoretical level but also on how LLMs can deal with hope across diverse linguistic and cultural contexts (Krafft et al., 2023).

References

Pranjal Aggarwal, Pasupuleti Chandana, Jagrut Nemade, Shubham Sharma, Sunil Saumya, and Shankar Biradar. 2023. Hope speech detection on social media platform. In Pradeep Kumar Roy and Asis Kumar Tripathy, editors, *Cybercrime in Social Media: Theory and Solutions*, 1st edition, pages 67–84. CRC Press.

AI@Meta. 2024. Llama 3 model card.

Pranav Anand and Valentine Hacquard. 2013. Epistemics and attitudes. *Semantics and Pragmatics*, 6:1–59.

Fazlourrahman Balouchzahi, Sabur Butt, Grigori Sidorov, and Alexander Gelbukh. 2023a. Reddit: Regret detection and domain identification from text. *Expert Systems with Applications*, 225:120099.

Fazlourrahman Balouchzahi, Grigori Sidorov, and Alexander Gelbukh. 2023b. Polyhope: Two-level hope speech detection from tweets. *Expert Systems with Applications*, 225:120078.

Claudia Bloeser and Titus Stahl. 2022. Hope. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Summer 2022 edition. Metaphysics Research Lab, Stanford University.

Michael Bratman. 1987. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press, Cambridge.

Cohan Sujay Carlos and Madhulika Yalamanchi. 2012. Intention analysis for sales, marketing and customer service. In *Proceedings of COLING 2012: Demonstration Papers*, pages 33–40, Mumbai, India. The COLING 2012 Organizing Committee.

Bharathi R. Chakravarthi, B. Bharathi, Joephine Griffith, Kalika Bali, and Paul Buitelaar, editors. 2023. *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*. INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria.

Bharathi Raja Chakravarthi. 2020. Hopeedi: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53.

Bharathi Raja Chakravarthi, B Bharathi, John P McCrae, Manel Zarrouk, Kalika Bali, and Paul Buitelaar, editors. 2022a. *LPS@LT-EDI-ACL2022:An Ensemble Approach about Hope Speech Detection*. Association for Computational Linguistics, Dublin, Ireland.

Bharathi Raja Chakravarthi, B Bharathi, John P McCrae, Manel Zarrouk, Kalika Bali, and Paul Buitelaar, editors. 2022b. *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics, Dublin, Ireland.

Bharathi Raja Chakravarthi, John P. McCrae, Manel Zarrouk, Kalika Bali, and Paul Buitelaar, editors. 2021. *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics, Kyiv.

George Chang, Han-Shen Huang, and Jane Hsu. 2013. Detecting chinese wish messages in social media: An empirical study. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):677–680.

Philip R. Cohen and Hector J. Levesque. 1990. Intention is choice with commitment. *Artificial Intelligence*, 42(2-3):213–261.

Cleo Condoravdi. 2002. Temporal interpretation of modals: Modals for the present and for the past. In Kaufmann B. Clark D. Beaver, S and L. Casillas, editors, *The Construction of Meaning*, pages 59–88. CSLI Publications.

Cleo Condoravdi and Sven Lauer. 2012. Imperatives: Meaning and illocutionary force. *Empirical issues in syntax and semantics*, 9:37–58.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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.

Li Dong, Furu Wei, Yajuan Duan, Xiaohua Liu, Ming Zhou, and Ke Xu. 2013. The automated acquisition of suggestions from tweets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27(1):239–245.

Jaklin Eliott and Ian Olver. 2002. The discursive properties of "hope": A qualitative analysis of cancer patients' speech. *Qualitative health research*, 12(2):173–193.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. 1999. The belief-desire-intention model of agency. In *Intelligent Agents V: Agents Theories, Architectures, and Languages: 5th International Workshop, ATAL'98 Paris, France, July 4–7, 1998 Proceedings 5*, pages 1–10. Springer.

Joseph John Godfrey. 2012. *A philosophy of human hope*, volume 9. Springer Science & Business Media.

Andrew B. Goldberg, Nathanael Fillmore, David Andrzejewski, Zhiting Xu, Bryan Gibson, and Xiaojin Zhu. 2009. May all your wishes come true: A study of wishes and how to recognize them. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, pages 263–271. Association for Computational Linguistics.

Thomas Grano. 2017. The logic of intention reports. *Journal of Semantics*, 34(4):587–632.

Patrick G. Grosz. 2012. *On the Grammar of Optative Constructions*. John Benjamins Publishing Company.

Sainan Guo. 2023. *Dynamic Adjusting Method on Emergency Plans for Group Decision-making Based on Regret Theory*. Ph.D. thesis, University of Vienna, Vienna. Doctor of Philosophy (PhD) in Economics/Business Administration: Management.

R. Abigail Hall. 2015. Drones: Public Interest, Public Choice, and the Expansion of Unmanned Aerial Vehicles. *Peace Economics, Peace Science, and Public Policy*, 21(2):273–300.

Charles L. Hamblin. 1973. Questions in Montague English. *Foundations of Language*, 10(1):41–53.

John S. Hammond, Ralph L. Keeney, and Howard Raiffa. 1998. The hidden traps in decision making. *Harvard Business Review*, 76(5):47–passim.

Irene Heim. 1992. Presupposition projection and the semantics of attitude verbs. *Journal of semantics*, 9(3):183–221.

Chadi Helwe, Tom Calamai, Pierre-Henri Paris, Chloé Clavel, and Fabian Suchanek. 2024. MAFALDA: A benchmark and comprehensive study of fallacy detection and classification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4810–4845, Mexico City, Mexico. Association for Computational Linguistics.

Albert O. Hirschman. 1970. *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Harvard University Press, Cambridge, Massachusetts.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jinsoo Hwang, Hyun Kim, and Woohyoung Kim. 2019. Investigating motivated consumer innovativeness in the context of drone food delivery services. *Journal of Hospitality and Tourism Management*, 38:102–110.

Ao Jia, Yu He, Yazhou Zhang, Sagar Uprety, Dawei Song, and Christina Lioma. 2022. Beyond emotion: A multi-modal dataset for human desire understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1512–1522, Seattle, United States. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Andreas M Krafft, Tharina Guse, and Alena Slezackova. 2023. *Hope across cultures: Lessons from the international hope barometer*. Springer Nature.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Anh Ngo and Hanh T. H. Tran. 2023. Zootopi at HOPE2023iberLEF: Is zero-shot chatgpt the future of hope speech detection? In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023)*. CEUR-WS.org.

Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. 2020. Hope speech detection: A computational analysis of the voice of peace. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 1881–1889.

Paul Portner. 2009a. *Modality*. Oxford Surveys in Semantics and Pragmatics. Oxford University Press, Oxford, New York.

Paul Portner. 2009b. *Modality*, volume 1. Oxford University Press.

Paul Portner. 2018. *Mood*. Oxford University Press.

Paul Portner and Barbara Partee. 2008. *Formal semantics: The essential readings*. John Wiley & Sons.

J. Ramanand, Krishna Bhavsar, and Niranjan Pedanekar. 2010. Wishful thinking - finding suggestions and 'buy' wishes from product reviews. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 54–61, Los Angeles, CA. Association for Computational Linguistics.

Aynat Rubinstein, Hillary Harner, Elizabeth Krawczyk, Daniel Simonson, Graham Katz, and Paul Portner. 2013. Toward fine-grained annotation of modality in text. In *Proceedings of the IWCS 2013 Workshop on Annotation of Modal Meanings in Natural Language (WAMM)*, pages 38–46.

Ted Ruffman, Lance Slade, Kate Rowlandson, Charlotte Rumsey, and Alan Garnham. 2003. How language relates to belief, desire, and emotion understanding. *Cognitive Development*, 18(2):139–158.

Roser Saurí and James Pustejovsky. 2009. Factbank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227–268.

Grigori Sidorov, Fazlourrahman Balouchzahi, Sabur Butt, and Alexander Gelbukh. 2023. Regret and hope on transformers: An analysis of transformers on regret and hope speech detection datasets. *Applied Sciences*, 13(6).

C. Richard Snyder. 2000. *Handbook of hope: Theory, measures, and applications*. Academic press.

Luis Alfonso Ureña López, Rafael Valencia García, Salud M. Jiménez Zafra, Miguel Ángel García Cumbreras, Daniel García Baena, Jos´ Antonio García Díaz, and Bharathi Raja Chakravarthi. 2023. Overview of HOPE at iberLEF 2023: Multilingual hope speech detection. *Procesamiento del Lenguaje Natural*, 71:371–381.

Elisabeth Villalta. 2008. Mood and gradability: An investigation of the subjunctive mood in spanish. *Linguistics and philosophy*, 31:467–522.

Bo Xu, Longjiao Li, Wei Luo, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2024. Beyond linguistic cues: Fine-grained conversational emotion recognition via belief-desire modelling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2318–2328, Torino, Italia. ELRA and ICCL.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. 2023. Starling-7b: Improving llm helpfulness & harmlessness with rlaif.

Yue Zhu. 2023. I2c-huelva at hope2023iberlef: Simple use of transformers for automatic hope speech detection. In *In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Lan- guage Processing (SEPLN 2023)*.

A Keywords to collect the HopeDrone corpus

The following keywords have been used to scrap data from Twitter[8] [9] and Reddit[10] about drones in the HopeDrone dataset: *drone, drone delivery, drone surveillance, drone taxi, military drone, agricultural drone, recreational drone, surveillance drone, reconnaissance drone, autonomous delivery drone, Amazon delivery drone, unmanned aerial*

[8] https://developer.twitter.com/en/docs/twitter-api, in conjunction with a developer account

[9] This dataset has been collected before the acquisition and subsequent re-branding of Twitter to X.

[10] https://www.reddit.com/dev/api/

vehicle, UAV, drone no-fly zone, drone laws, and drone regulation.

B The CDB Annotation Campaign

B.1 Confusion Matrix in the CDB annotations

Figure 5 shows the confusion matrix between our two annotators after the double annotation of 2,000 instances in a subset of the WISH and PolyHope datasets following the CDB scheme (NO DECISION cases have been removed).

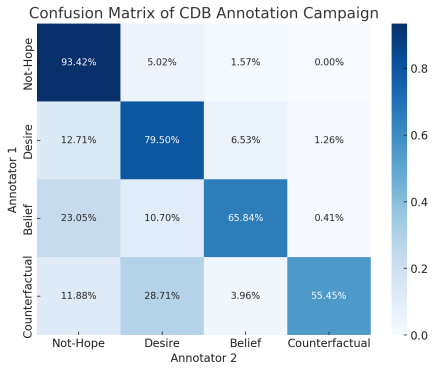

Figure 5: Confusion matrix between annotators in the CDB dataset.

B.2 Disagreement Cases

Common Ambiguities for COUNTERFACTUALS. Some common ambiguities arise from combining attitude verbs with other verbs in the past tense. In example (13), although *knew* is the past tense of *know*, the overall meaning of the sentence is centered in the present and is characterized as a DESIRE.

(13) *I wish I knew more people in the Lansing area, I need to find hoop runs*

(14) *I wish Cali had more of those. Get them sometimes from mountains with monsoons but usually spark fires*

(15) *Oh awe I was hoping to see you there but we will have to hang out then while im out there for KCON. I'm waiting for concerts as well but I just couldn't miss out on KCON. I have to see the boyz since the last two shows were cancelled*

(16) *he should not be fired this professor what he said was on his own spare time he did not talk about it in class*

Example (14) provides a clear illustration of the tense structure of a true COUNTERFACTUAL, featuring a complete past construction. The speaker expresses regret over the fact that California is different from their expectations. Similarly, in example (15), despite differences in grammatical structure, the overall meaning conveys regret. Therefore, both examples (14) and (15) should be classified as instances of COUNTERFACTUAL.

Misclassifying COUNTERFACTUAL as BELIEF is also possible. In example (16), the presence of *should* adds complexity because it implies a sense of duty or obligation (deontology). However, the context suggests an expectation that was not fulfilled, and the speaker is expressing regret about the actions of others. Example (16) should therefore be classified as COUNTERFACTUAL.

Common Ambiguities for DESIRE. The past orientation at the beginning of (17) is not sufficient to categorize it as COUNTERFACTUAL. Despite referencing problems in the past, the speaker is expressing hope that the same will not happen to them.

(17) *well thats about it and as people said in earlier previews that the headfone jack gets messed up stil havent happened to me and i hope it doesnt*

Although the attitude verb in example (18) is in the past tense, it conveys the speaker's mere desire to have an option other than an iPod at some point in the present or future.

(18) *not to mention that everyone else at the office has an ipod so i was kinda hoping for a good alternative*

While sentences like (19) and (20) may initially appear as mere desires, it is crucial to consider the epistemic connections conveyed by the speaker. In example (19), there is more than a mere wish; there is a forecast that the lawsuits will result in a consequence: exposing hypocrisy. Moving to example (20), it is rooted in the deontic aspect of the expectation. Here, the speaker is not merely expressing a desire but is forecasting the future based on their moral beliefs. For both these examples, the **Belief** category is the most appropriate label, and we will spell out their precise function and meaning in the next section.

(19) *As a Catholic I am hopeful these lawsuits are coming to show the hypocrisy of the Supreme Court.*

(20) *to make amends i think the bush twins jenna the fat one and barbara the one that looks like a moose should go to baghdad*

Finally, misclassifying NOT-HOPE as a DESIRE can appear straightforward, as in example (21), where an imperative structure is present. Imperatives often imply planning for others and might be classified as DESIRE. However, in this instance, understanding the overall context of the sentence reveals that the speaker is simply offering high praise for a product. Similarly, in example (22), instead of an imperative, we face a question, further demonstrating the need to grasp the broader meaning beyond the surface structure. In both examples (21) and (22), we should use NOT-HOPE as the label.

(21) *what do you want me to do go over and kiss the camera*

(22) *where did you get your m d i just want to know*

Common Ambiguities for BELIEF. In some cases, desire may contain elements of belief. In example (23), the speaker merely justifies their hope regarding their cousin testifying. While there is a deontic expectation expressed in *Italy's honor*, it is subordinate and used only to justify their hope. Similarly, in example (24), the speaker, a supporter of Barcelona Futbol Club, attempts to justify their hopes for the team's future. There is an illusory construction of epistemic authority, but the speaker acknowledges they are only guessing, as evidenced by *it wasn't what we thought* in the past. Expressions such as *have to* and *I'm hopeful* help configure the example as a DESIRE rather than a BELIEF. In both examples (23) and (24), the DESIRE category would be more appropriate.

(23) *I'm hoping and praying that my cousin, Pat Cipollone (Pasquale Cipollone ("Chip-alone")) will testify tomorrow. He must testify for his, his parents', his family, and everyone that came from Gallo Matese in Italy's honor.*

(24) *Nah, it wasn't what we thought. Hopefully we'll get Frenkie done in the next 2-3 days. Barca have to sell him before 1st July, so I'm hopeful. I also hope we get Eriksen to join us.*

Common Ambiguities for NOT-HOPE. However, the mere presence of some attitude verbs is insufficient to categorize an instance into one of the hope-related categories. For instance, in example (25), despite the presence of *want* and *#hopeful*, the instance does not appear to constitute an expression of desire; rather, it seems more like a mere elaboration or irony from the speaker. Example (26) appears to be a straightforward question, and example (27) is another elaboration about a past expectation that was not accomplished.

(25) *I always wonder if the opposition voters just don't want to be called out in their small communities. Rather just privately vote. #hopeful*

(26) *how do we actually anticipate moves like the one that happened last week?*

(27) *thinking the cd was faulty i went to the linksys website hoping to simply download the setup program but it s not available*

C Experimental Settings

C.1 Train-Test Split

Figure 6 shows the distribution of CDB classes as well as datasets (among WISH, PolyHope, HopeEDI and HopeDrone) in the train/test sets.

C.2 Description of the Models

We make use of the following open source models:

LLAMA3$_{8B}$ (AI@Meta, 2024), developed by Meta, is pre-trained on 500 billion tokens from diverse sources, including books, articles, and web data. The model employs mixed-precision training and gradient checkpointing to manage computational efficiency. Performance benchmarks indicate an F1 score of 76.4 on SQuAD. This model presents a knowledge cuttof up to March of 2023

LLAMA3$_{70B}$ (AI@Meta, 2024), provided by Meta and pretrained on 1.5 trillion tokens. Performance benchmarks indicate an F1 score of 85.6 on SQuAD, highlighting its robust capabilities in text comprehension. This model presents a knowledge cuttof up to December of 2023.

MISTRAL$_{7B\text{ - INSTRUCT - V0.3}}$ (Jiang et al., 2023) by Mistralai is pre-trained on 600 billion tokens, including instructional manuals, academic texts, and technical documentation. Fine-tuning is performed on a curated set of instructional datasets, employing techniques like curriculum learning and knowledge distillation. The model scores 88.7 on the ITE (Instructional Text Evaluation) benchmark and demonstrates superior performance in the clarity and coherence of instructional text.

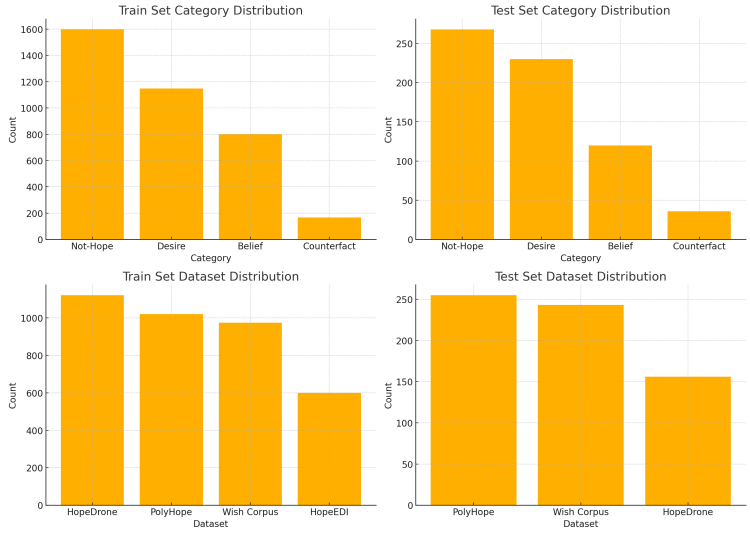

Figure 6: Distribution of categories and corpora in our CDB train/test sets, presented in a decreasing order of frequency.

Aquila2$_{\text{7B}}$ (Beijing Academy of Artificial Intelligence, BAAI)[11]is pre-trained on 1.4 trillion tokens from diverse sources, including literature and web content. The model utilizes techniques like masked language modeling and sequence-to-sequence learning. Aquila2 achieves a BLEU score of 27.5 on the WMT English-German translation task and an F1 score of 90.3 on SQuAD 2.0, showcasing its strong performance in translation and question-answering.

Starling-LM $_{\text{7B-beta}}$ (Zhu et al., 2023) by Nexusflow (Berkeley Artificial Intelligence Research) is pre-trained on 800 billion tokens from sources like chat logs, forums, and social media. Fine-tuning involves datasets such as ConvAI and DailyDialog, employing masked language modeling and next-sentence prediction. The model achieves a BLEU score of 25.3 on the ConvAI benchmark and an F1 score of 89.1 on Persona-Chat, reflecting its high performance in dialogue management.

Vicuna $_{\text{13B-v1.5}}$ (Zheng et al., 2023), developed by Large Model Systems Organization (LMSYS Org), is pre-trained on a trillion tokens from a mixture of web crawls, academic papers, and dialogue datasets. Fine-tuning includes RLHF (Reinforcement Learning from Human Feedback) to optimize conversational quality. The model achieves a performance of 86.5 on the MMLU benchmark and 92.3 on the HELM benchmark. Advanced techniques like dynamic batching and parallel training are used to enhance efficiency and scalability.

C.3 Models Parameters

All transformer models (i.e., Bert and RoBerta) have been trained with the AdamW optimizer, a fixed batch size of 32, a learning rate of $2e-5$ for BERT, $5e-5$ (resp. $1e-5$) for binary (resp. multiclass) classification with RoBERTa.

LLMs hyper-parameters are shown in Table 8. To reduce the verbosity and keep the models focused on our scheme, we set the temperature to 0.001.

Parameter	Value
num_return_sequences	1
top_p	0.1
max_new_tokens (baseline)	5
max_new_tokens (CARP)	200
temperature	0.001
do_sample	True

Table 8: Parameters used for testing LLM models.

For fine-tuning Mistral with LoRA, we used a learning rate of $1e-4$ and warmup ratio of 0.1. We fine-tuned the model for 5 epoches, with batch size equal to 32.

C.4 Prompts

We present all the prompts used in our experiments in Figure 7.

[11] https://github.com/FlagAI-Open/Aquila2

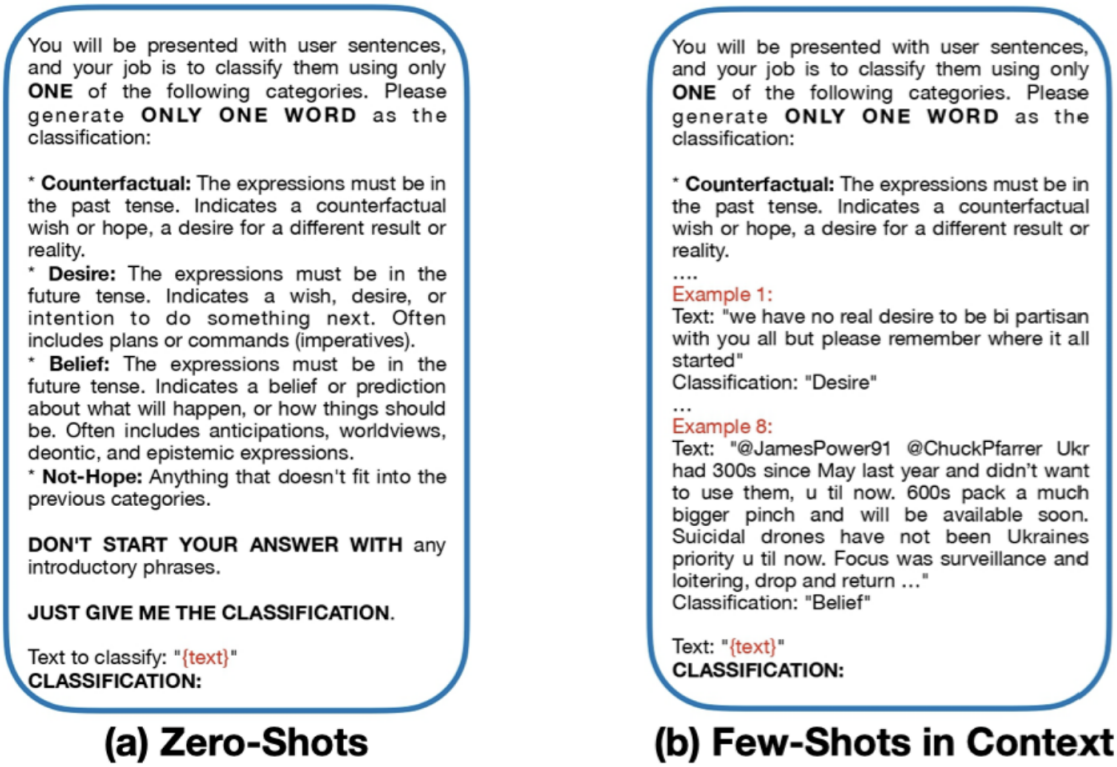

Figure 7: Prompts utilized to test the LLMs. Bold, red, and underline were used only to highlight differences and patterns in the configurations.