ANNOTATION GUIDELINE

Phase 1: 5 annotators labeling 3400 comments should strictly follow the guidelines below

index	token	norm	is_norm	certainty
index: 110				
0	cận	cẩn	#	
1	thận	thận		
2	nhà	nha	#	
3	ae	anh em	#	?

Annotation Interface

1. In column ‘norm’, if there is no need for normalization, the token remains unchanged; otherwise, it is altered. It should be noted that certain tokens, like links, email addresses, icons, emoticons, tokens written in languages other than Vietnamese, or common abbreviations of proper noun are regarded as non-standard but do not require normalization.
2. In column ‘is_norm’, add '#' if normalization is needed, otherwise, leave the cell empty.
3. Label column ‘certainty’ with "?". Leave it blank if the annotator is unsure of how to normalize.
4. Tokens are all written with lowercase letters, even with proper nouns.
5. If there are multiple standard forms, list them in the same cell, separated by commas.
6. The numerical digits remain the same (1 person, 3 chickens), while the non-quantitative digits are translated into words such as '1 số' → 'một số' (some), 'chị 3' → 'chị ba' (the third sister).
7. Exclamatory words like 'hahaaa' and 'huhuu' remain unchanged.
8. If the tokens are foreign words written in Vietnamese transliteration, such as 'biu ty phun' (beautiful), they should be normalized by adding a dash '-' between syllables ('biu ty phun' - 'biu-ty-phun') or corrected into Vietnamese if there is an appropriate word.
9. If a token is split into multiple cells, the annotator should put the word normalized in the first cell and leave the remaining cells blank.

Index: 330				
0	cây	k-icm	#	
1	a		#	
2	tê		#	
3	em		#	

10. Normalize Vietnamese tone, for instance, 'hoá' →'hóa' (transform), 'nguỵ' → 'ngụy' (fake).

Phase 2: Integrate the labels by different annotators

1. When there is disagreement in normalization, if one person is show certainty (without '?') while another are uncertain (with '?'), the final standard form is assigned by the former

sent_idx	token_idx	token	norm1	norm2	is_norm1	is_norm2	certainty1	certainty2	norm
20	0	bamtp	bạn	bamtp	#			?	bạn
20	1	chạy	chạy	chạy					chạy
20	2	ngay	ngay	ngay					ngay
20	3	ddi	đi	đi	#	#			đi
20	4	:))	:))	:))					:))

2. When there is disagreement in normalization, if both annotators show certainty or uncertainty, another annotator who has not join phase 1 will label those cases.

sent_idx	token_idx	token	norm1	norm2	is_norm1	is_norm2	certainty1	certainty2	norm
2984	0	thầy	thầy	thầy					thầy
2984	1	có	có	có					có
2984	2	hướng	hướng	hướng					hướng
2984	3	nghiệp	nghiệp	nghiệp					nghiệp
2984	4	nghề	nghề	nghề					nghề
2984	5	này	này	này					này
2984	6	hông	không	không	#	#			không
2984	7	zạ	dạ	vậy	#	#			dạ

Note: After the annotation stage:

1. Drop comments with no non-standard tokens

sent_idx	token_idx	token	norm1	norm2	is_norm1	is_norm2	certainty1	certainty2	norm
3004	0	cực	cực	cực					cực
3004	1	kì	kì	kì					kì
3004	2	thuyết	thuyết	thuyết					thuyết
3004	3	phục	phục	phục					phục
3004	4	:(((	:(((	:(((					:(((

2. Duplicate sentence with multiple modifications

non-standard	standard	replicate1	replicate2	replicate3
150k	150000, 150 nghìn, 150 ngàn	150000	150 nghìn	150 ngàn
2	2	3	4	5
chai	chai	chai	chai	chai
đó	đó	đó	đó	đó

Dataset's size: 3400 comments → 2181 comments