An In-depth Analysis of Implicit and Subtle Hate Speech Messages

Nicolas Benjamin Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata

Université Côte d'Azur, Inria, CNRS, I3S, France

HS, Implicitness and Subtlety

Explicit HS is **easily identifiable** (words whose definition is hateful), whereas **Implicit HS** employs exaggeration, metaphor, irony, etc., **coding** a **message's** true **nature** for ML models. **Subtle HS** depends on user's perception **pushing out the attention** from the hateful meaning.

Implicit HS: *"I'm either in North Florida or Nigeria sometimes i can't tell the difference"*

Subtle HS: *"As a brit my knowledge of american law is somewhat lacking but even i know that this holder groid has committed treason."*

Objectives and Challenges

Security in **social media** has grown substantially due to the **enforcement** of **anti-hate speech** (HS) policies. The most relevant approaches to **detect HS messages** rely on **supervised ML architectures**. But current state-of-the-art (SOTA) models **detect** well **explicit** HS messages but **fail** on **implicit** and **subtle** ones.

Our contributions:

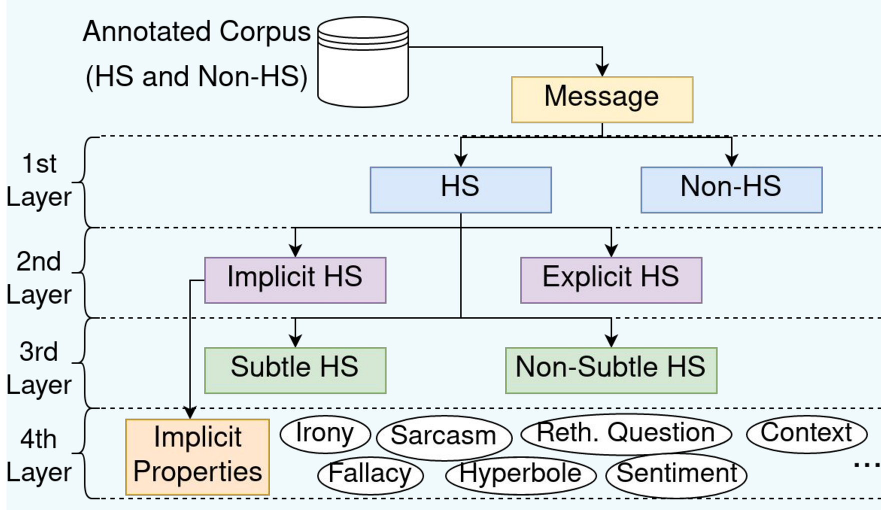
Provide a newly created **HS dataset** named **ISHate** with **implicit** and **subtle labels**, fine-grained **implicit properties** annotations, and **augmentation methods**.

Classify messages with **SOTA** models on **two 3-label supervised tasks**, Task A (Non-HS/Explicit HS/Implicit HS) and Task B (Non-HS/Non-Subtle HS/Subtle HS).

The ISHate Dataset

We use 7 datasets, already annotated with HS and Non-HS labels (1st layer), collected from **users** potentially **prone to produce HS content**.

On top of HS messages, **implicit** and **subtle labels** are added (2nd, 3rd layers). Implicit cases are fine-grained annotated with **18 linguistic features** (4th layer).



	Train		Dev		Test	
Label	#	%	#	%	#	%
Non-HS	12508	.614	2680	.614	2681	.614
Explicit HS	7007	.344	1501	.344	1501	.344
Implicit HS	866	.042	186	.043	186	.043
Non-HS	12508	.614	2680	.614	2681	.614
Non-Subtle	7691	.377	1648	.377	1648	.377
Subtle	182	.009	39	.009	39	.009

ISHate label schema and data distribution of each layer.

Experiments and Results

We propose the following HS SOTA models:

- Universal Sentence Encoder (USE)+SVM.
- Bert-like models: Vanilla BERT, HateBERT, and DeBERTa.

We augment the target minority classes of both tasks, **Implicit HS** and **Subtle HS**:

- Replace Scalar Adverbs (RSA)
- Add Adverbs to Verbs (AAV)
- Replace Named Entities (RNE)
- Replace In-Domain Expressions (RI)
- Replace Adjectives (RA)
- Easy Data Augmentation (EDA)
- Back Translation (BT)
- Generative Models (GM)
- Generative Models with human valid. (GM+R.)
- Use all methods (ALL)

Label	RSA	AAV	RNE	RI	RA	EDA	BT	GM	GM + R.	ALL
Implicit HS	6848	7032	828	817	467	6935	748	200	82	23957
Subtle HS	3192	3136	480	210	172	2912	179	200	204	10685

Label	ORIG	RSA	AAV	RNE	RI	RA	EDA	BT	GM	GM + R.	ALL
Non-HS	.614	.459	.456	.59	.590	.600	.458	.592	.608	.611	.282
Explicit HS	.344	.257	.256	.33	.331	.336	.257	.332	.340	.342	.158
Implicit HS	.042	.283	.288	.08	.079	.064	.286	.076	.052	.046	.560
Non-HS	.614	.531	.532	.600	.607	.609	.537	.608	.608	.608	.403
Non-Subtle	.377	.326	.327	.369	.374	.374	.330	.374	.374	.374	.248
Subtle	.009	.143	.141	.032	.019	.017	.133	.018	.019	.019	.350

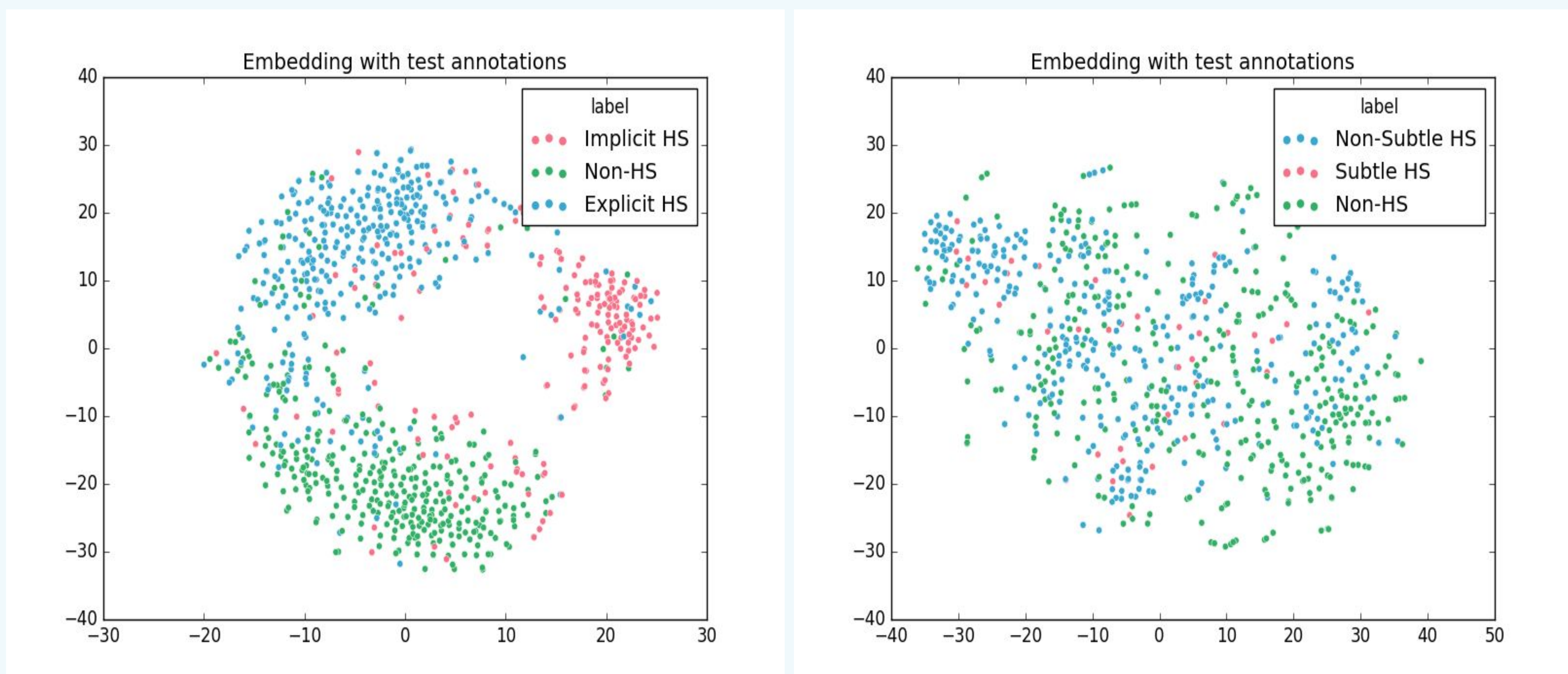
Number of additional implicit/subtle messages and distribution (%) generated by each augmentation method.

The **best performing models** on Task A and B are **HateBERT+ALL**, and **USE+SVM+BT**, respectively.

The most **misclassified implicit properties** in task A are **Inference (53%)**, **Context (41%)**, **Sentiment (40%)**, **Exaggeration (24%)**, and **Extralinguistic knowledge (24%)**. In task B **word order** and **circumlocution** affect models' performances.

Model	Non-HS			Explicit HS			Implicit HS			Non-HS			Non-Subtle HS			Subtle HS		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
USE+SVM	.888	.866	.877	.766	.803	.784	.399	.382	.390	.891	.868	.879	.783	.832	.807	.667	.103	.178
BERT	.903	.893	.898	.81	.833	.821	.394	.371	.382	.902	.891	.897	.819	.846	.832	.250	.103	.145
HateBERT	.904	.89	.897	.811	.849	.829	.447	.382	.412	.903	.890	.897	.814	.850	.831	.143	.026	.043
DeBERTa	.927	.899	.913	.825	.880	.851	.467	.419	.442	.920	.893	.906	.823	.877	.849	.375	.077	.128
HateBERT+ALL	**.903**	**.896**	**.899**	**.827**	**.827**	**.827**	**.502**	**.559**	**.529**	.903	.881	.892	.816	.844	.830	.391	.462	.424
BERT+BT	.909	.887	.898	.824	.826	.825	.459	.608	.523	.898	.900	.899	.839	.832	.835	.304	.359	.329
DeBERTa + BT	.919	.885	.902	.830	.857	.844	.428	.543	.479	.920	.897	.908	.835	.876	.855	.385	.256	.308
USE+SVM+BT	.897	.856	.876	.782	.787	.785	.403	.645	.496	**.892**	**.868**	**.880**	**.789**	**.831**	**.809**	**.739**	**.436**	**.548**
BERT+RNE	.897	.897	.897	.807	.829	.818	.455	.349	.395	.899	.895	.897	.826	.839	.833	.400	.256	.312
DeBERTa+RI	.922	.894	.908	.821	.878	.849	.460	.398	.427	.910	.894	.902	.828	.860	.843	.364	.205	.262
HateBERT+GM	.901	.898	.899	.824	.827	.825	.414	.425	.419	.899	.898	.899	.831	.834	.832	.250	.231	.240
HateBERT+GM+R	.905	.891	.898	.816	.835	.826	.408	.419	.414	.894	.898	.896	.826	.826	.826	.192	.128	.154

Relevant results of SOTA models on tasks A and B.



t-SNE Embedding of HateBERT+ALL and USE+SVM+BT in the test sets of task A and B respectively.