Comment	I feel embarrassed by the fact that I even clicked on this video	This is probably the worst song ever released.	It sounds like a donkey who's high on skittles	Possibly the worst thing I've ever heard. And seen.	She may or may not be a Jew but she 's certainly stupid , she seems to think the Blacks wo n't kill her alongside every other White they can get their dirty hands on , what a muppet !	Does it surprise you that jews promote faggotry ?	You are stupid. This is a bad answer	I just came across this piece of garbage .	Black men used to be slaves	I'm looking to study at a college in South Africa later in my life .	Imagine disliking a 3 hour course for free. Smh...	The only course that I actually sit down and finish. No fluff, informative and the small humor comments make it binge worthy. Thank you!!	i feel guilty for not paying for this course.... its that good...	There is no ads. This is too generous	I hate homosexuals	I love being homosexual	you must be fucking kidding!!	this is fucking amazing!	don't tell me bullshit,	holy shit! the best song ever!	Me hace mucha gracia lo patético que eres. Das bastante pena.	Nadie confia en ti, porque eres un mentiroso	no entiendo como puedes existir, eres un ser despreciable	hay que exterminar a algunos seres humanos	hoy en dia los politicos son todos igual de asquerosos	odio a los homosexuales	amo a los homosexuales	los nazis conquistaron gran parte de europa hace menos de un siglo	necesitamos que los nazis vuelvan y acaben con las ratas	joder, es la hostia!	me cago en la puta, a ver cuando vuelve a sacar una nueva canción...	me he comprado estos cereales y están joddidamente buenos
Comment Description	Ironic criticism, not that toxic	Strong criticism, a little bit toxic	toxic, insult	A little bit toxic (would depend on the context)	severely toxic, identity attack, profanity, insult	toxic, insult, profanity, identity attack	toxic, insult	toxic results. Some models also detect insult	Not toxic. Can be biased identiti attack. The best results (>0.5) IBM	Not toxic					toxic, identity attack perspectiveapi detects a lot of labels	not toxic. can be biased perspectiveapi biased (around 0.5)	not toxic, but with bad words				toxic, insult	insult	toxic, insult eapi detoxify doesn't detect toxicity	toxic, threat	maybe toxic	toxic, identity attack	can be biased perspectiveapi: identity attack (biased)	not toxic, could be biased (mentions "nazis") perspectiveapi: identity attack (biased)	toxic, threat, identity attack	not toxic, but contain bad words		
Results Analysis, Incoherences...	Not classified as toxic	IBM and original, the highest toxic ratio	tensorflow detects < 0.3 ibm doesn't detect insult	Perspectiveapi the highest, but < 0.5	Mostly well rated	The most confident: perspectiveapi	Well rated			Not toxic							perspectiveapi, tensorflow: detect insult ibm: detects identity attack			best: perspectiveapi	best: perspectiveapi	best: perspectiveapi	detoxify: not toxic perspectiveapi: toxic	well rated		detoxify doesn't detect the toxicity	detoxify: considers toxicity perspectiveapi: detects profanity, also toxicity		
Detoxify – original																															
TOXICITY	0.0007	0.4942	0.6252	0.2672	0.9889	0.8505	0.9837	0.7907	0.8693	0.0007	0.0025	0.0013	0.0008	0.0006	0.9656	0.5890	0.9936	0.9847	0.9819	0.9650											
SEVERE_TOXICITY	0.0001	0.0004	0.0008	0.0004	0.2056	0.0604	0.0223	0.0015	0.0419	0.0001	0.0001	0.0001	0.0001	0.0001	0.0031	0.0036	0.1520	0.1032	0.0518	0.0364											
IDENTITY_ATTACK	0.0001	0.0007	0.0058	0.0010	0.8497	0.0075	0.0018	0.7063	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001	0.7913	0.0776	0.0035	0.0029	0.0021	0.0023											
INSULT	0.0002	0.0238	0.1621	0.0055	0.8795	0.5765	0.9253	0.1190	0.2854	0.0002	0.0002	0.0002	0.0002	0.0002	0.2334	0.0070	0.3811	0.1365	0.1497	0.0771											
THREAT	0.0001	0.0006	0.0007	0.0009	0.1311	0.0115	0.0013	0.0005	0.0091	0.0001	0.0001	0.0001	0.0001	0.0001	0.0337	0.0079	0.0032	0.0035	0.0025	0.0022											
OBSCENE	0.0002	0.0101	0.0131	0.0049	0.7749	0.5256	0.6231	0.0928	0.2023	0.0002	0.0002	0.0002	0.0002	0.0002	0.1830	0.0157	0.9795	0.9674	0.9383	0.9062											
Detoxify – unbiased																															
TOXICITY	0.0012	0.0421	0.9208	0.0321	0.9945	0.6608	0.9872	0.9961	0.7727	0.0010	0.0004	0.0004	0.0004	0.0005	0.0004	0.9031	0.1511	0.8858	0.9241	0.9705	0.9528										
SEVERE_TOXICITY	0.0000	0.0000	0.0000	0.0000	0.0063	0.0015	0.0002	0.0001	0.0007	0.0000	0.0000	0.0000	0.0000	0.0000	0.0026	0.0004	0.0002	0.0004	0.0008	0.0015											
IDENTITY_ATTACK	0.0000	0.0003	0.0004	0.0004	0.7059	0.5636	0.0031	0.0012	0.7165	0.0001	0.0000	0.0000	0.0000	0.0000	0.8103	0.1177	0.0027	0.0045	0.0075	0.0102											
INSULT	0.0004	0.0219	0.8869	0.0131	0.9864	0.1466	0.9963	0.9946	0.1560	0.0002	0.0001	0.0001	0.0001	0.0001	0.2487	0.0096	0.0735	0.0850	0.3557	0.1468											
THREAT	0.0000	0.0003	0.0003	0.0006	0.0111	0.0056	0.0014	0.0006	0.0074	0.0000	0.0000	0.0000	0.0000	0.0000	0.0762	0.0029	0.0063	0.0067	0.0116	0.0187											
SEXUALLY_EXPLICIT	0.0000	0.0001	0.0002	0.0001	0.0033	0.0084	0.0014	0.0006	0.0025	0.0000	0.0000	0.0000	0.0000	0.0000	0.0153	0.0087	0.0136	0.0221	0.0187	0.0219											
OBSCENE	0.0000	0.0007	0.0015	0.0006	0.0428	0.0495	0.0014	0.0013	0.0036	0.0000	0.0000	0.0000	0.0000	0.0000	0.0044	0.0025	0.9304	0.9665	0.9843	0.9865											
Detoxify – multilingual																															
TOXICITY	0.0049	0.0312	0.5646	0.0397	0.9902	0.5264	0.9925	0.9492	0.6532	0.0006	0.0146	0.0005	0.0006	0.0012	0.9079	0.3173	0.9911	0.8474	0.9336	0.8864	0.4969	0.6985	0.0702	0.6303	0.0336	0.8977	0.4802	0.2147	0.3358	0.9886	0.9764	0.8990
Tensorflow.js																															
TOXICITY	0.0007	0.0207	0.2972	0.0124	0.9731	0.8766	0.9959	0.9817	0.7441	0.0004	0.0018	0.0003	0.0004	0.0003	0.8899	0.3179	0.8895	0.7445	0.9388	0.9187											
SEVERE_TOXICITY	0.0000	0.0000	0.0000	0.0000	0.0030	0.0010	0.0000	0.0000	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0025	0.0001	0.0001	0.0003	0.0001	0.0006											
IDENTITY_ATTACK	0.0000	0.0004	0.0007	0.0001	0.5055	0.6284	0.0085	0.0055	0.7228	0.0001	0.0000	0.0000	0.0000	0.0000	0.7612	0.2014	0.0072	0.0148	0.0085	0.0284											
INSULT	0.0002	0.0143	0.2269	0.0070	0.9550	0.3565	0.9956	0.9808	0.0882	0.0001	0.0005	0.0001	0.0002	0.0002	0.2659	0.0112	0.3732	0.1039	0.6999	0.5623											
THREAT	0.0001	0.0002	0.0005	0.0002	0.0589	0.0649	0.0015	0.0025	0.0276	0.0001	0.0001	0.0000	0.0001	0.0000	0.0877	0.0124	0.0071	0.0118	0.0051	0.0172											
SEXUALLY_EXPLICIT	0.0001	0.0001	0.0005	0.0001	0.0250	0.0241	0.0011	0.0008	0.0061	0.0000	0.0001	0.0000	0.0000	0.0000	0.0367	0.0095	0.0556	0.0727	0.0428	0.0833											
OBSCENE	0.0001	0.0003	0.0011	0.0005	0.1490	0.0291	0.0061	0.0032	0.0125	0.0000	0.0002	0.0000	0.0001	0.0000	0.0499	0.0032	0.9012	0.8150	0.8974	0.9093											
PerspectiveAPI																															
TOXICITY	0.3276	0.2426	0.6364	0.4033	0.8576	0.8785	0.9322	0.8585	0.7944	0.0896	0.0895	0.1526	0.0878	0.0792	0.9210	0.4400	0.9324	0.8059	0.9297	0.9023	0.9491	0.8706	0.8860	0.8341	0.8880	0.9560	0.5111	0.2694	0.9035	0.5322	0.8706	0.5821
SEVERE_TOXICITY	0.1514	0.0526	0.2929	0.1552	0.7783	0.8081	0.6269	0.4827	0.6452	0.0535	0.0389	0.0696	0.0326	0.0253	0.7665	0.2925	0.8081	0.5122	0.6054	0.6252	0.5306	0.2025	0.5306	0.9419	0.5204	0.9108	0.3521	0.2037	0.9419	0.1447	0.7322	0.1311
IDENTITY_ATTACK	0.1342	0.0625	0.2788	0.2982	0.9222	0.9424	0.2768	0.1526	0.8922	0.1011	0.0969	0.0548	0.0393	0.0386	0.9681	0.5852	0.2391	0.1268	0.2857	0.3094	0.1524	0.1149	0.2720	0.7535	0.6693	0.9986	0.8610	0.6672	0.9254	0.0611	0.3187	0.0966
INSULT	0.2198	0.2056	0.7143	0.4041	0.8925	0.8982	0.9344	0.8955	0.6467	0.0447	0.0602	0.1760	0.0450	0.0471	0.7692	0.2632	0.7912	0.2946	0.6607	0.6356	0.9786	0.9995	0.9140	0.3878	0.9241	0.8579	0.3049	0.2294	0.8619	0.1571	0.7069	0.2284
PROFANITY	0.2825	0.1064	0.4413	0.1590	0.8421	0.8899	0.8437	0.6222	0.5128	0.0832	0.0414	0.0482	0.0299	0.0394	0.7348	0.4304	0.9560	0.9559	0.9722	0.9693	0.2938	0.1694	0.3592	0.0939	0.3592	0.7430	0.2782	0.0333	0.5204	0.9806	0.9918	0.9651
THREAT	0.2463	0.0770	0.1460	0.1141	0.7739	0.2988	0.1820	0.3154	0.2959	0.1241	0.1071	0.2152	0.0920	0.0575	0.5258	0.2138	0.2971	0.1380	0.1163	0.2452	0.1868	0.1355	0.3758	0.9913	0.1868	0.5937	0.3019	0.2421	0.9913	0.2423	0.3019	0.0702
SEXUALLY_EXPLICIT	0.2668	0.0533	0.1973	0.0709	0.3789	0.5361	0.0866	0.1171	0.1227	0.1383	0.0881	0.0621	0.0523	0.0725	0.2969	0.6282	0.4322	0.1893	0.2207	0.1145											
FLIRTATION	0.4066	0.3134	0.3251	0.3126	0.5609	0.3904	0.1988	0.3317	0.1993	0.3463	0.4007	0.4156	0.2474	0.2724	0.1877	0.6688	0.5375	0.4056	0.2027	0.3211											
(NYT) ATTACK_ON_AUTHOR	0.4525	0.2061	0.9637	0.2393	0.1228	0.4139	0.6853	0.8522	0.0108	0.0443	0.2079	0.7131	0.1884	0.0690	0.0812	0.1184	0.1716	0.1338	0.0489	0.0347											
(NYT) ATTACK_ON_COMMENTER	0.0428	0.0830	0.2644	0.3981	0.1285	0.7804	0.9726	0.1590	0.2151	0.0757	0.7267	0.3063	0.2983	0.0447	0.3396	0.2714	0.7276	0.1221	0.0910	0.0934											
(NYT) INCOHERENT	0.4891	0.3369	0.6689	0.3666	0.6085	0.4350	0.1311	0.1851	0.4512	0.4613	0.4796	0.4758	0.5778	0.6085	0.4473	0.6656	0.0526	0.1226	0.1024	0.0459											
(NYT) INFLAMATORY	0.0944	0.1637	0.5881	0.2727	0.9089	0.7910	0.3721	0.4800	0.7232	0.2736	0.2449	0.1935	0.2178	0.0921	0.7279	0.6304	0.3563	0.3545	0.2784	0.2719											
(NYT) LIKELY_TO_REJECT	0.7708	0.3430	0.8344	0.3864	0.9786	0.9669	0.9914	0.9204	0.9045	0.3870	0.7645	0.5854	0.4929	0.6653	0.9508	0.9046	0.9997	0.9994	0.9986	0.9995											
(NYT) OBSCENE	0.0441	0.0701	0.5854	0.0588	0.5854	0.1180	0.0551	0.0704	0.0259	0.0180	0.1316	0.1134	0.1069	0.0573	0.0419	0.2524	0.8881	0.9892	0.9904	0.9937											
(NYT) SPAM	0.1433	0.0189	0.3554	0.0408	0.0714	0.0161	0.0045	0.0126	0.0107	0.1268	0.0463	0.3403	0.0881	0.3440	0.0259	0.2044	0.0087	0.1742	0.0935	0.0425											
(NYT) UNSUBSTANTIAL	0.8551	0.8579	0.8308	0.8660	0.3096	0.5956	0.6224	0.9208	0.8327	0.7727	0.8579	0.7747	0.8308	0.8321	0.8577	0.8623	0.1709	0.2605	0.4679	0.1296											
IBM MAX																															
TOXICITY	0.0003	0.7772	0.9519	0.2296	0.9887	0.8385	0.9914	0.9307	0.5668	0.0003	0.0003	0.0003	0.0003	0.0003	0.9856	0.4354	0.9811	0.9448	0.9605	0.8928											
SEVERE_TOXICITY	0.0001	0.0003	0.0007	0.0004	0.0559	0.0343	0.0417	0.0007	0.0038	0.0003	0.0001	0.0001	0.0001	0.0001	0.0961	0.0007	0.0283	0.0447	0.0055	0.0043											
IDENTITY_ATTACK	0.0001	0.0003	0.0022	0.0007	0.9346	0.6133	0.0134	0.0008	0.5218	0.0003	0.0001	0.0001	0.0001	0.0001	0.2594	0.0020	0.9380	0.9239	0.6965	0.6033											
INSULT	0.0002	0.0023	0.0679	0.0009	0.8556	0.1637	0.9772	0.0925	0.2202	0.0002	0.0001	0.0002	0.0002	0.0002	0.0668	0.0011	0.0032	0.0072	0.0023	0.0013											
THREAT	0.0001	0.0022	0.0037	0.0051	0.0218	0.0015	0.0021	0.0004	0.0012	0.0002	0.0001	0.0001	0.0001	0.0001	0.6559	0.0038	0.0375	0.0177	0.0095	0.0054											
OBSCENE	0.0002	0.0018	0.0013	0.0007	0.2983	0.0659	0.7458	0.0365	0.0345	0.0002	0.0001	0.0002	0.0002	0.0001	0.9667	0.0371	0.0012	0.0021	0.0007	0.0005											