** SCRIPT START in TEST_TRAIN mode **

READING IN DATA

READING EXCEL DATA

XLS_DF:

	TITLE	...	CAT_7
0	The English language in the Philippines	...	0.0
1	Creole and "Cajan"	...	0.0
2	Aesop in Negro Dialect	...	0.0
3	The Chinook Jargon	...	0.0
4	Indian Songs and English Verse	...	0.0
..	...	...	...
268	Continuity versus English Influence in the Wes...	...	0.0
269	Unstressed been: Past and Present in African A...	...	0.0
270	“Put the Groceries Up”: Comparing Black and Wh...	...	0.0
271	“My Presiden(t) and Firs(t) Lady Were Black”:S...	...	0.0
272	“You’re So Not Going to Believe This”:The Use ...	...	0.0

[273 rows x 25 columns]

> XLS_ROWS PREVIEW:

>> {'TITLE': 'The English language in the Philippines', 'AUTHORS': 'Yule, E.S', 'DOI': '10.2307/452557', 'INFO_COMPLETE': 'Y', 'PDF_DOWNLOADED': 'Y', 'MACHINE_READABLE': 'Y', 'PDF_NAME': '10-2307_452557', 'VENUE': 'American Speech', 'YEAR': 1925, 'MONTH': 'NOV', 'VOLUME': 1, 'ISSUE': 2, 'UNKNOWN': nan, 'CAT_0': nan, 'CAT_1': 0.0, 'CAT_2': 0.0, 'CAT_2_1': 0.0, 'CAT_3': 0.0, 'CAT_3_1': nan, 'CAT_4': 0.0, 'CAT_4_1': nan, 'CAT_5': 0.0, 'CAT_5_1': 1.0, 'CAT_6': 0.0, 'CAT_7': 0.0}

>> {'TITLE': 'Creole and "Cajan"', 'AUTHORS': 'William A. Read', 'DOI': 'https://www.jstor.org/stable/452755', 'INFO_COMPLETE': 'Y', 'PDF_DOWNLOADED': 'Y', 'MACHINE_READABLE': '?', 'PDF_NAME': 'jstor_452755', 'VENUE': 'American Speech', 'YEAR': 1926, 'MONTH': 'JUN', 'VOLUME': 1, 'ISSUE': 9, 'UNKNOWN': 1.0, 'CAT_0': nan, 'CAT_1': 0.0, 'CAT_2': 0.0, 'CAT_2_1': 0.0, 'CAT_3': 0.0, 'CAT_3_1': nan, 'CAT_4': 0.0, 'CAT_4_1': nan, 'CAT_5': 0.0, 'CAT_5_1': 0.0, 'CAT_6': 0.0, 'CAT_7': 0.0}

>> ..

>> {'TITLE': '“My Presiden(t) and Firs(t) Lady Were Black”:Style, Context, and Coronal Stop Deletion in the Speech of Barack and Michelle Obama', 'AUTHORS': 'Nicole Holliday', 'DOI': '10.1215/00031283-6903954', 'INFO_COMPLETE': 'Y', 'PDF_DOWNLOADED': 'Y', 'MACHINE_READABLE': nan, 'PDF_NAME': '10-1215_00031283-6903954', 'VENUE': 'American Speech', 'YEAR': 2017, 'MONTH': 'Winter', 'VOLUME': 92, 'ISSUE': 4, 'UNKNOWN': nan, 'CAT_0': nan, 'CAT_1': 1.0, 'CAT_2': 0.0, 'CAT_2_1': 0.0, 'CAT_3': 0.0, 'CAT_3_1': nan, 'CAT_4': 0.0, 'CAT_4_1': nan, 'CAT_5': 0.0, 'CAT_5_1': 0.0, 'CAT_6': 0.0, 'CAT_7': 0.0}

>> {'TITLE': '“You’re So Not Going to Believe This”:The Use of Genx so in Constructions with Future going to in American English', 'AUTHORS': 'Ulrike Stange', 'DOI': '10.1215/00031283-4395168', 'INFO_COMPLETE': 'Y', 'PDF_DOWNLOADED': 'Y', 'MACHINE_READABLE': 'Y', 'PDF_NAME': '10-1215_00031283-4395168', 'VENUE': 'American Speech', 'YEAR': 2017, 'MONTH': 'Winter', 'VOLUME': 92, 'ISSUE': 4, 'UNKNOWN': nan, 'CAT_0': nan, 'CAT_1': 1.0, 'CAT_2': 0.0, 'CAT_2_1': 0.0, 'CAT_3': 0.0, 'CAT_3_1': nan, 'CAT_4': 0.0, 'CAT_4_1': nan, 'CAT_5': 0.0, 'CAT_5_1': 0.0, 'CAT_6': 0.0, 'CAT_7': 0.0}

>> found and read XLSX file (american-speech-dataset-complete-rows-final-sep-13-2023.xlsx - 273 rows)

📄 READING PDF DATA
>> found 273 PDF files -- 2791761 total words (avg: 10226)

> PDF TEXT PREVIEW (after preprocessing):
everybody likes pizza doesnt he or she author(s): george jochnowitz source: american speech autumn NUM vol. NUM no. CIT pp. NUM NUM published by: duke university press stable url: https://www.jstor.org/stable/NUM references linked references are available on jstor for this article: https://www.jstor.org/stable/NUM seq=NUM &cid=pdf reference#references_tab_contents you may need to log in to jstor to access the linked references. jstor is a notforprofit service that helps scholars researchers and students discover use and build upon a wide range of content in a trusted digital archive. we use information technology and tools to increase productivity and facilitate new forms of scholarship. for more information about jstor please contact support@jstor.org. your use of the jstor archive indicates your acceptance of the terms & conditions of use available at https://about.jstor.org/terms duke university press is collaborating with jstor to digitize preserve and extend access to american speech this content downloaded from NUM .NUM .NUM .NUM on wed NUM aug NUM :NUM :NUM +NUM :NUM all use subject to

https://about.jstor.org/terms everybody likes pizza doesnt he or she george jochnowitz college of staten island w hile reading margaret mitchells gone with the wind my younger daughter found what she called a grammatical error in the text of the NUM edition: everyone was very polite and kind to her because he felt sorry for her.. . CIT. when i decided to write an essay arguing that they is grammatical with indefinite antecedents i checked the text of the CIT edition. this time the passage read everyone was very polite and kind to her because they felt sorry for her... CIT. apparently between NUM and NUM some anonymous copy editor decided to correct mitchells prose and in doing so created the jarring phrase that had offended my daughter. what makes the NUM version of the sentence sound so unnatural j. j. CIT says that if the singularcongruent form im me ..

CONSTRUCTING DOCUMENT OBJECTS

>> Completed making 273 document objects - two quick preview documents:

PRINT_OUT for document <10-2307_454860.pdf>, titled <Everybody Likes Pizza, Doesn't He or She?> (authors: George Jochnowitz)
DOI 10.2307/454860 | YEAR 1982 | MONTH Autumn | VOLUME 57 | ISSUE 3
CATEGORY LIST: ['CAT_7']

TEXT: everybody likes pizza doesnt he or she author(s): george jochnowitz source: american speech autumn NUM vol. NUM no. CIT pp. NUM NUM published by: duke university press stable url: https://www.jstor ..

X_VECTOR: []
Y_GOLDS: []
Y_PREDS: [-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1]

PRINT_OUT for document <10-1215_00031283-78-2-171.pdf>, titled <REVISITING THE CREOLIST HYPOTHESIS: COPULA VARIABILITY IN GULLAH AND SOUTHERN RURAL AAVE> (authors: TRACEY L. WELDON)
DOI 10.1215/00031283-78-2-171 | YEAR 2003 | MONTH Summer | VOLUME 78 | ISSUE 2
CATEGORY LIST: ['CAT_1']

TEXT: american speech vol. NUM no. NUM summer NUM copyright © NUM by the american dialect society NUM revisiting the creolist hypothesis: copula variability in gullah and southern rural aave tracey l. we ..

X_VECTOR: []
Y_GOLDS: []
Y_PREDS: [-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1]

>> Converted ALL_DOCS into XLSX, saved @ XLSX/all_docs.xlsx ..

📄 MAKING and FIT-TRANSFORMING TF-IDF VECTORISER
>> X_RAW has been made (not displaying all texts, as that would be too long)

>> X_ALL is currently:
  (0, 10673) 0.007019822157239164
  (0, 7702) 0.013738075062721112
  (0, 4505) 0.010819075895447741
  (0, 4103) 0.006065570459905716
  (0, 12445) 0.009992461362039578
  (0, 28159) 0.008281242397481663
  (0, 22027) 0.005406781057638 28
  (0, 8578) 0.014519925584897087
  (0, 15577) 0.01715590744679743
  (0, 9698) 0.016260915543206113
  (0, 29242) 0.004523686090556611
  (0, 4340) 0.012594066685149303
  (0, 22589) 0.009695947093605274
  (0, 2557) 0.010207101297229447
  (0, 22367) 0.007635266967870832
  (0, 12291) 0.010437654541315483
  (0, 20099) 0.007059965231704932
  (0, 17399) 0.009890450744 21879
  (0, 9498) 0.014999495302963 61
  (0, 14797) 0.009603073178698211
  (0, 21893) 0.008163830534 81097
  (0, 8189) 0.013410293746115897
  (0, 21459) 0.01715590744679743
  (0, 23929) 0.003655629810689888
  (0, 28954) 0.003203232096494771 86
  : :
  (9, 18709) 0.026273924069796967
  (9, 14021) 0.03213706689122 12
  (9, 29323) 0.001064891562377 6714

(9, 12210) 0.026482763855353687
(9, 27974) 0.0012360316046511562
(9, 24781) 0.0022097318633611637
(9, 20525) 0.0019083831698267619
(9, 27770) 0.005683511873630298
(9, 7842) 0.0022846890495475626
(9, 3482) 0.01705053562089089 3
(9, 20986) 0.0011048659316805819
(9, 20294) 0.0012313538075945 59
(9, 4394) 0.079569166230824 16
(9, 17788) 0.012314275726198977
(9, 28479) 0.0009472519789383829
(9, 18136) 0.5503533997632005
(9, 2020) 0.0031423106921684 62
(9, 24625) 0.045468094989042 38
(9, 978) 0.018945039578767 66
(9, 24462) 0.0010493624936253226
(9, 1996) 0.0011423445247737813
(9, 23617) 0.009666441416744725
(9, 18674) 0.026620031954912 03
(9, 11626) 0.010966252071757035
(9, 7560) 0.0025034318136738 34 ..

FEATURE_NAMES preview: ['__' '___' '_____' '______' '_____'
'_____' '__adj' '_l' '_loc'
'_np' '_num' '_story' '_v' 'aa' 'aaal' 'aae' 'aal' 'aam' 'aan' 'aaron'
'aav' 'aave' 'aavegullah' 'aaves' 'aavespeaking' 'ab' 'aba' 'aback'
'aban' 'abandon' 'abandoned' 'abandoning' 'abandonment' 'abbott'
'abbrevi' 'abbreviate' 'abbreviated' 'abbreviation' 'abbreviations'
'abby' 'abc' 'abdomen' 'abe' 'aberdeen' 'aberdeens' 'aberrant'
'aberration' 'abiding' 'abigail' 'abilities' 'ability' 'able' 'ables'
'ablex' 'ably' 'abnaki' 'abnormal' 'abnormalities' 'abnormally' 'aboard'
'abolished' 'abolition' 'abolitionism' 'abolitionist' 'abolitionists'
'abooklength' 'aboriginal' 'aborigines' 'abortion' 'abound' 'abounded'
'abounding' 'abounds' 'about' 'aboutface' 'aboutlanguage' 'aboutthe'
'aboutwhether' 'above' 'abovementioned' 'abra' 'abraham' 'abrahams'
'abram' 'abre' 'abridged' 'abroad' 'abrupt' 'abruptly' 'absence'
'absencenum' 'absent' 'absentia' 'abso' 'absolute' 'absolutely'
'absolutelynecessary' 'absoluterelative' 'absolutes' 'absorbed'] ..

ENCODING LABELS with MLB

>> Y_RAW is currently:
[['CAT_7'], ['CAT_6'], ['CAT_1'], ['CAT_2_1'], ['CAT_2', 'CAT_3'], ['CAT_1', 'CAT_2'], ['CAT_1', 'CAT_2'], ['CAT_1'], ['CAT_6'], ['CAT_1', 'CAT_2']] ..

>> Y_ALL is currently:
[[0 0 0 0 0 0 0 0 1]
 [0 0 0 0 0 0 0 1 0]
 [1 0 0 0 0 0 0 0 0]
 [0 0 1 0 0 0 0 0 0]
 [0 1 0 1 0 0 0 0 0]
 [1 1 0 0 0 0 0 0 0]
 [1 1 0 0 0 0 0 0 0]
 [1 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 1 0]
 [1 1 0 0 0 0 0 0 0]] ..

SPLITTING DATA (TRAIN-TEST SPLIT using SKF)

/home/michaelinwords/.local/lib/python3.8/site-packages/sklearn/model_selection/_split.py:725: UserWarning: The least populated class in y has only 1 members, which is less than n_splits=5.
  warnings.warn(
>> After splitting, there are 218 training documents/rows and 55 test documents/rows -- 273 documents/rows total

TRAINING CLASSIFIER

X_TRAIN is:

  (0, 10673)	0.007019822157239164
  (0, 7702)	0.013738075062721112
  (0, 4505)	0.010819075895447741
  (0, 4103)	0.006065570459905716
  (0, 12445)	0.009992461362039578
  (0, 28159)	0.008281242397481663
  (0, 22027)	0.005406781057638 28
  (0, 8578)	0.014519925584897087
  (0, 15577)	0.01715590744679743
  (0, 9698)	0.016260915543206113
  (0, 29242)	0.004523686090556611

(0, 4340) 0.012594066685149303
(0, 22589) 0.009695947093605274
(0, 2557) 0.010207101297229447
(0, 22367) 0.007635266967870832
(0, 12291) 0.010437654541315483
(0, 20099) 0.007059965231704932
(0, 17399) 0.00989045074421879
(0, 9498) 0.014999495302963 61
(0, 14797) 0.009603073178698211
(0, 21893) 0.008163830534810 97
(0, 8189) 0.013410293746115897
(0, 21459) 0.01715590744679743
(0, 23929) 0.003655629810689888
(0, 28954) 0.003203232096494771 86
: :
(217, 26323) 0.37280033271247853
(217, 12696) 0.24965814136401 28
(217, 26787) 0.08097020800995 01
(217, 16139) 0.002614827303789449
(217, 19339) 0.037431588547251 26
(217, 26528) 0.038798224671434 42
(217, 9966) 0.06747517334162507
(217, 18526) 0.04470230233882661
(217, 2031) 0.000924135933946 2543
(217, 1530) 0.020316565072153764
(217, 21694) 0.001147798816423 1266
(217, 29323) 0.000948186759688 7503
(217, 20525) 0.005097712437505905
(217, 27770) 0.005904077667392194
(217, 3482) 0.049762940339448496
(217, 4394) 0.07169237167547665
(217, 17788) 0.010964715668014074
(217, 28479) 0.000843439666770313 4
(217, 18136) 0.5229325933975943
(217, 24625) 0.021929431336028 15
(217, 978) 0.026990069336650 03
(217, 24462) 0.000934359570234457 9
(217, 23617) 0.002151766454088419
(217, 18674) 0.015237423804115325
(217, 11626) 0.005326050626890849

Y_TRAIN is:

[[0 0 0 ... 0 0 1]
 [0 0 0 ... 0 1 0]
 [1 0 0 ... 0 0 0]
 ...
 [1 0 0 ... 0 0 0]
 [1 0 0 ... 0 0 0]
 [1 0 0 ... 0 0 0]]

REVIEWING PERFORMANCE

-- This training was on 218 documents (out of 273 total documents)

TFIDF settings - PREPROCESSOR None, NGRAM_RANGE (1, 1), BINARY False, ANALYSER word, MIN_DF 2
Stratified K-Fold - KFOLD_SPLITS: 5

>> Top 20 positive features for class 0 (AAL/AAVL (CAT_1)):
['black (1.31898)', 'negro (1.11219)', 'gullah (1.01165)', 'of (0.80596)', 'aave (0.78851)', 'cit (0.68906)', 'copula (0.65224)', 'african (0.56668)', 'bin (0.56081)', 'be (0.54722)', 'in (0.53839)', 'that (0.47093)', 'he (0.45112)', 'white (0.43413)', 'deletion (0.42131)', 'aae (0.38362)', 'bev (0.37931)', 'auxiliary (0.35687)', 'bees (0.35174)', 'data (0.34547)']

>> Top 20 negative features for class 0 (AAL/AAVL (CAT_1)):
['japanese (-0.86126)', 'spanish (-0.68715)', 'indian (-0.60056)', 'gay (-0.53812)', 'french (-0.42010)', 'women (-0.36373)', 'word (-0.36265)', 'fronting (-0.34323)', 'words (-0.33370)', 'or (-0.32853)', 'female (-0.31098)', 'indians (-0.30885)', 'male (-0.30535)', 'mexican (-0.30450)', 'men (-0.29782)', 'names (-0.29751)', 'korean (-0.29021)', 'items (-0.28763)', 'woman (-0.28219)', 'hawaiian (-0.28122)']

>> Top 20 positive features for class 1 (African Am. (CAT_2)):
['african (0.78214)', 'black (0.72830)', 'white (0.58709)', 'blacks (0.54589)', 'fronting (0.51688)', 'bees (0.49705)', 'yall (0.37624)', 'whites (0.36426)', 'bes (0.34904)', 'texana (0.33669)', 'racial (0.30904)', 'americans (0.30719)', 'afrikaans (0.30676)', 'american (0.29883)', 'dutch (0.29784)', 'negro (0.29255)', 'benum (0.29124)', 'weren (0.28997)', 'divergence (0.28646)', 'south (0.28159)']

>> Top 20 negative features for class 1 (African Am. (CAT_2)):
['english (-0.63262)', 'is (-0.52910)', 'the (-0.48486)', 'japanese (-0.41579)', 'spanish (-0.29933)', 'gay (-0.28510)', 'jstor (-0.28228)', 'indian (-0.27176)', 'to (-0.25311)', 'gullah (-0.24298)', 'creole (-0.24022)', 'he (-0.23513)', 'it (-0.23272)', 'pidgin (-0.21929)', 'yo (-0.20759)', 'adj (-0.19781)', 'language (-0.19104)', 'de (-0.19079)', 'bin (-0.17735)', 'french (-0.17514)']

>> Top 20 positive features for class 2 (African Diaspora (CAT_2_1)):
['creole (0.79551)', 'mi (0.55362)', 'jamaican (0.51889)', 'jamaica (0.44919)', 'is (0.39320)', 'afrikaans (0.37110)', 'caribbean (0.36176)', 'dutch (0.33745)', 'barbados (0.33498)', 'na (0.33197)', 'guiana (0.31261)', 'di (0.30093)', 'west (0.28272)', 'slaves (0.26262)', 'surinam (0.26091)', 'dem (0.25837)', 'cho (0.25275)', 'barbadian (0.24492)', 'it (0.24012)', 'bev (0.23888)']

>> Top 20 negative features for class 2 (African Diaspora (CAT_2_1)):
['of (-0.79859)', 'num (-0.36763)', 'japanese (-0.29510)', 'names (-0.27247)', 'and (-0.26624)', 'american (-0.21361)', 'the (-0.19549)', 'are (-0.18877)', 'gay (-0.15698)', 'speakers (-0.15472)', 'aave (-0.14975)', 'was (-0.14316)', 'http (-0.13065)', 'univ (-0.13040)', 'for (-0.12605)', 'our (-0.12462)', 'edu (-0.12250)', 'spanish (-0.11825)', 'bulletnum (-0.11662)', 'dukeupress (-0.11528)']

>> Top 20 positive features for class 3 (Mexican Am. & Latinx (CAT_3)):
['spanish (1.48136)', 'mexican (0.75269)', 'sp (0.58392)', 'chicano (0.53728)', 'devoicing (0.46308)', 'hispanic (0.41980)', 'english (0.36813)', 'puerto (0.36810)', 'durham (0.29162)', 'spanishspeaking (0.29052)', 'mexicanamerican (0.28733)', 'rico (0.26394)', 'quotative (0.25400)', 'mexicans (0.24816)', 'el (0.24398)', 'cog (0.22965)', 'french (0.22258)', 'mexico (0.21940)', 'speakers (0.21776)', 'bilingual (0.20801)']

>> Top 20 negative features for class 3 (Mexican Am. & Latinx (CAT_3)):
['the (-0.67115)', 'num (-0.58004)', 'of (-0.40743)', 'and (-0.40266)', 'to (-0.29146)', 'in (-0.28133)', 'black (-0.27125)', 'japanese (-0.26917)', 'negro (-0.24505)', 'for (-0.22685)', 'names (-0.22262)', 'on (-0.20834)', 'from (-0.19504)', 'he (-0.17820)', 'is (-0.17775)', 'indian (-0.16559)', 'white (-0.16522)', 'cit (-0.16471)', 'gullah (-0.16308)', 'univ (-0.15736)']

>> Top 20 positive features for class 4 (Native Am. (CAT_4)):

['indian (1.42504)', 'lumbee (0.85380)', 'indians (0.71062)', 'french (0.66764)', 'colorado (0.55713)', 'cherokee (0.51379)', 'louisiana (0.47722)', 'names (0.40808)', 'river (0.35219)', 'lumbees (0.32842)', 'tribe (0.30918)', 'robeson (0.28932)', 'ccr (0.27525)', 'papago (0.27355)', 'songs (0.26717)', 'missouri (0.26186)', 'name (0.25851)', 'coeur (0.25473)', 'chinook (0.25383)', 'county (0.25236)']

>> Top 20 negative features for class 4 (Native Am. (CAT_4)):
['num (-1.70006)', 'japanese (-0.42169)', 'negro (-0.38659)', 'black (-0.36365)', 'that (-0.35461)', 'african (-0.33156)', 'creole (-0.30210)', 'to (-0.25206)', 'in (-0.24508)', 'gullah (-0.24209)', 'gay (-0.22171)', 'cit (-0.20607)', 'speech (-0.18635)', 'adj (-0.18441)', 'as (-0.18101)', 'speakers (-0.16934)', 'bin (-0.15958)', 'male (-0.15929)', 'sp (-0.15632)', 'you (-0.15564)']

>> Top 20 positive features for class 5 (Asian Am./Pacific Is. (CAT_5)):
['hawaiian (0.83664)', 'hawaii (0.67148)', 'japanese (0.49067)', 'chinese (0.46004)', 'names (0.45988)', 'nisei (0.39563)', 'asian (0.33719)', 'colorado (0.32318)', 'korean (0.29600)', 'name (0.17214)', 'honolulu (0.16768)', 'pidgin (0.13654)', 'fri (0.13534)', 'surnames (0.13451)', 'instead (0.13367)', 'hawaiians (0.13179)', 'ka (0.12908)', 'americans (0.12229)', 'bergen (0.11864)', 'characters (0.11734)']

>> Top 20 negative features for class 5 (Asian Am./Pacific Is. (CAT_5)):
['num (-0.72869)', 'that (-0.46221)', 'and (-0.40039)', 'in (-0.40032)', 'the (-0.39320)', 'cit (-0.37724)', 'of (-0.31453)', 'to (-0.31023)', 'black (-0.20002)', 'negro (-0.18012)', 'african (-0.16879)', 'this (-0.16136)', 'he (-0.15867)', 'on (-0.15200)', 'from (-0.14355)', 'with (-0.13726)', 'was (-0.13454)', 'creole (-0.13124)', 'not (-0.12469)', 'aave (-0.11971)']

>> Top 20 positive features for class 6 (Asian Diaspora (CAT_5_1)):
['japanese (1.88440)', 'japan (0.71287)', 'bamboo (0.61858)', 'adj (0.58182)', 'items (0.56516)', 'english (0.56147)', 'korean (0.47284)', 'pidgin (0.47250)', 'vb (0.41058)', 'borrowings (0.36975)', 'malay (0.31392)', 'chinese (0.29458)', 'tea (0.28698)', 'nouns (0.27324)', 'missionaries (0.26419)', 'taksan (0.25405)', 'dictionaries (0.24657)', 'gi (0.23883)', 'words (0.23768)', 'loanwords (0.23629)']

>> Top 20 negative features for class 6 (Asian Diaspora (CAT_5_1)):
['the (-0.81467)', 'num (-0.69456)', 'of (-0.65702)', 'in (-0.46889)', 'cit (-0.36412)', 'and (-0.26408)', 'that (-0.26135)', 'african (-0.23894)', 'black (-0.23492)', 'negro (-0.21117)', 'creole (-0.19690)', 'univ (-0.18632)', 'http (-0.16879)', 'speech (-0.16788)', 'white (-0.16077)', 'names (-0.15496)', 'edu (-0.15495)', 'articlepdf (-0.15194)', 'dukeupress (-0.15194)', 'americanspeech (-0.15194)']

>> Top 20 positive features for class 7 (Women's Language (CAT_6)):
['lady (0.79165)', 'women (0.77088)', 'woman (0.58189)', 'female (0.56893)', 'male (0.52875)', 'gender (0.51451)', 'sexuality (0.46696)', 'gay (0.45816)', 'sex (0.38322)', 'gentleman (0.35438)', 'womens (0.34422)', 'men (0.32962)', 'aeu (0.32215)', 'responses (0.31619)', 'color (0.31192)', 'sexual (0.30988)', 'pitch (0.29963)', 'penis (0.29554)', 'ms (0.27782)', 'age (0.26364)']

>> Top 20 negative features for class 7 (Women's Language (CAT_6)):
['english (-0.71031)', 'of (-0.69818)', 'the (-0.61472)', 'is (-0.47341)', 'in (-0.43805)', 'african (-0.39537)', 'black (-0.32925)', 'creole (-0.29414)', 'american (-0.29351)', 'negro (-0.28014)', 'japanese (-0.27974)', 'spanish (-0.26278)', 'from (-0.25774)', 'names (-0.25723)', 'he (-0.25215)', 'aave (-0.22896)', 'be (-0.20404)', 'gullah (-0.19529)', 'indian (-0.19474)', 'de (-0.18517)']

>> Top 20 positive features for class 8 (LGBTQ Speech (CAT_7)):
['gay (1.24370)', 'closet (0.71005)', 'yo (0.55557)', 'sexuality (0.48736)', 'lesbian (0.45167)', 'lesbians (0.41198)', 'sexual (0.38606)', 'munson (0.35150)', 'pitch (0.30042)', 'generic (0.25148)', 'gender (0.25025)', 'pronoun (0.24264)', 'masculine (0.21726)', 'regan (0.21085)', 'participants (0.20457)', 'aspnum (0.20420)', 'singular (0.20214)', 'orientation (0.19891)', 'israeli (0.19355)', 'http (0.19280)']

>> Top 20 negative features for class 8 (LGBTQ Speech (CAT_7)):
['the (-1.16358)', 'num (-0.93727)', 'in (-0.60505)', 'of (-0.48978)', 'english (-0.33214)', 'and (-0.32614)', 'from (-0.21702)', 'jstor (-0.21506)', 'all (-0.19388)', 'african (-0.17927)', 'is (-0.17838)', 'to (-0.17826)', 'black (-0.17659)', 'american (-0.17358)', 'for (-0.17312)', 'names (-0.16732)', 'be (-0.16207)', 'https (-0.15767)', 'org (-0.15693)', 'negro (-0.15044)']

SUBSET ACCURACY: 9.09%
CLASSIFICATION REPORT:

	precision	recall	f1-score	support
AAL/AAVL (CAT_1)	1.00	0.18	0.31	22
African Am. (CAT_2)	1.00	0.00	0.00	9
African Diaspora (CAT_2_1)	1.00	0.00	0.00	5
Mexican Am. & Latinx (CAT_3)	1.00	0.00	0.00	4
Native Am. (CAT_4)	1.00	0.00	0.00	8

Asian Am./Pacific Is. (CAT_5)	1.00	0.00	0.00	2
Asian Diaspora (CAT_5_1)	1.00	0.00	0.00	5
Women's Language (CAT_6)	1.00	0.00	0.00	5
LGBTQ Speech (CAT_7)	1.00	0.00	0.00	3
micro avg	1.00	0.06	0.12	63
macro avg	1.00	0.02	0.03	63
weighted avg	1.00	0.06	0.11	63
samples avg	1.00	0.09	0.09	63

per_label_accuracy: 0.88

📄 SAVING MODEL
>> Successfully saved classifier as model.joblib and vectoriser as vectoriser.joblib

📄 SCRIPT (train mode) COMPLETED SUCCESSFULLY ✅