Data pre-processing

Chunk size = 50000 ; No.of chunks = 20

Total data loaded: 2.706532 GB

Total size of all tokens on disk: 1.799708 GB

Label: Neutral, News Type: Neutral

Label: Reliable, News Type: Neutral

Label: Fake, News Type: Neutral

Removed Rows per Chunk:

Chunk 1:

Removed rows: 2874

URLs count: 36693

Dates count: 11088

Numeric values count: 600285

fake: 33092

reliable: 11282

neutral: 2752

Chunk 2:

Removed rows: 1153

URLs count: 23841

Dates count: 22969

Numeric values count: 655155

fake: 28996

reliable: 15902

neutral: 3949

Chunk 3:

Removed rows: 3244

URLs count: 6738

Dates count: 21555

Numeric values count: 620212

fake: 6902

reliable: 27871

neutral: 11983

Chunk 4:

Removed rows: 1219

URLs count: 9381

Dates count: 13788

Numeric values count: 565945

fake: 17159

reliable: 25885

neutral: 5737

Chunk 5:

Removed rows: 850

URLs count: 5302

Dates count: 3516

Numeric values count: 420857

fake: 37459

reliable: 7560

neutral: 4131

Chunk 6:

Removed rows: 4250

URLs count: 7377

Dates count: 5531

Numeric values count: 439838

fake: 33943

reliable: 3718

neutral: 8089

Chunk 7:

Removed rows: 5347

URLs count: 3901

Dates count: 6662

Numeric values count: 308490

fake: 40203

reliable: 2852

neutral: 1598

Chunk 8:

Removed rows: 5612

URLs count: 13361

Dates count: 4755

Numeric values count: 512868

fake: 36606

reliable: 3827

neutral: 3955

Chunk 9:

Removed rows: 9084

URLs count: 6821

Dates count: 2582

Numeric values count: 379573

fake: 20632

reliable: 15213

neutral: 5071

Chunk 10:

Removed rows: 919

URLs count: 3955

Dates count: 7589

Numeric values count: 461294

fake: 15769

reliable: 17701

neutral: 15611

Chunk 11:

Removed rows: 307

URLs count: 8001

Dates count: 10391

Numeric values count: 487691

fake: 24140

reliable: 16882

neutral: 8671

Chunk 12:

Removed rows: 1851

URLs count: 9318

Dates count: 6325

Numeric values count: 404251

fake: 17552

reliable: 22998

neutral: 7599

Chunk 13:

Removed rows: 1293

URLs count: 10842

Dates count: 10880

Numeric values count: 566648

fake: 10081

reliable: 33805

neutral: 4821

Chunk 14:

Removed rows: 142

URLs count: 24300

Dates count: 12536

Numeric values count: 761147

fake: 35441

reliable: 13823

neutral: 594

Chunk 15:

Removed rows: 346

URLs count: 9303

Dates count: 13053

Numeric values count: 595397

fake: 13256

reliable: 32704

neutral: 3694

Chunk 16:

Removed rows: 557

URLs count: 6059

Dates count: 8379

Numeric values count: 591267

fake: 12903

reliable: 24690

neutral: 11850

Chunk 17:

Removed rows: 636

URLs count: 2673

Dates count: 2649

Numeric values count: 549789

fake: 31431

reliable: 7650

neutral: 10283

Chunk 18:

Removed rows: 524

URLs count: 2832

Dates count: 1886

Numeric values count: 629055

fake: 37240

reliable: 6359

neutral: 5877

Chunk 19:

Removed rows: 608

URLs count: 531

Dates count: 852

Numeric values count: 547293

fake: 42130

reliable: 3437

neutral: 3825

Chunk 20:

Removed rows: 185

URLs count: 273

Dates count: 668

Numeric values count: 467496

fake: 31164

reliable: 1914

neutral: 16737

Total Vocab Size: 233278916

Total Vocab Size after stemming: 156504197

Top 100 most frequent words (excluding numeric values) in the whole dataset:

one: 1104920

use: 1082356

would: 1043978

state: 988491

peopl: 915718

time: 844398

like: 825500

said: 815384

year: 810756

system: 789822

also: 766704

us: 763082

new: 753008

make: 718628

trump: 704033

go: 678856

get: 612597

govern: 604668

american: 583594

even: 578922

say: 575663

presid: 570687

report: 553234

tor: 543281

right: 543228

rec: 542376

obama: 540147

call: 517047

come: 516502

nation: 515149

see: 508933

could: 504736

mani: 501727

work: 492294

first: 483540

may: 471591

day: 463516

world: 456379

two: 452649

take: 450265

know: 446063

think: 439997

tail: 439791

commun: 436395

need: 434306

want: 433407

polit: 430409

public: 429360

countri: 425025

last: 420439

support: 414456

way: 413742

democrat: 407865

war: 389591

oper: 386133

well: 385736

republican: 376537

back: 368335

secur: 365386

much: 362716

news: 361499

look: 359797

hous: 357569

includ: 355997

thing: 352467

unit: 343230

help: 340054

order: 339977

good: 334230

group: 330374

parti: 326985

live: 320996

law: 319165

show: 313021

elect: 312557

chang: 308397

power: 306485

vote: 305582

made: 302513

next: 301341

end: 300170

week: 299518

offici: 296437

sinc: 296264

point: 294448

america: 293558

follow: 290922

issu: 289928

fact: 287943

gener: 287346

white: 286316

internet: 286269

continu: 285203

forc: 282890

plan: 280462

still: 278725

part: 277537

media: 277149

million: 276672

believ: 275381