Leung Wai Liu

JPMorgan AI Research

August 9, 2022

Twitter Relabelling Dataset Observations

**TRAINING DATASET**

10 - #fridaynightlights – no break

11 - #harrypotter marathon – hashtag then a word

252 - #napflow – is actually a brand

263 – early bird special – phrase

267 - #campuschat – no break

293 – oem window mount – phrase

295 – Monday nov – phrase

301 – fitness business – phrase

330 – prime minister – phrase

334 - #mnd charity – hashtag then a word

339/340 - #hawethornen\*\*\*\*\* #noelton – two consecutive hashtags broken up

417 - #24audiolaunch – no break

423 - #soundcloud – brand

456 - #cityvreal - ?

501 – zig zag – phrase

506 - #studentexperience – no break

508 - #uniexperience – no break

512 – guilty pear – break

515/516 - #amazon #affiliate – two consecutive hashtags broken up

538 – yella Habibi – phrase

544 – po ako – phrase

555/556/557 - #socialmediaqueen #choicefemalehottie #choiceinstagrammer – three consecutive hashtags broken up

562 – asif Siddiqui daily – phrase

569 – marathon filibuster – phrase

578 - #gameofthrones s5 – hashtag then a word

591 - #designshanghai – no-break

593 – rugby league – phrase

619 - #contestalert – no break

620 - #sultantraileronsony – no break (4 words)

623 - #educatorsplaybook – no break

631 – caravan vl – phrase

633 - #200daysmsg2 – no break

637 - #genstemgirls – no break

638 – nike sb – phrase

639 – Wednesday apr-6-2016 – phrase

641 - #rednoseforkids – no break

653 – hongera kenya7s – phrase

660 – u8/10/12/14/16/19 – one word, but mislabaled as an entire phrase?

700/701/702 - #ralphlauren #tods #originalpenguin – 3 consecutive hashtags broken up

796 – prime minister – phrase

804 – foam roller – phrase

825 – teen choice – phrase

833 – lady gaga – phrase

838 - #demsinphilly – no break

889 - #choiceanteencipatedmovieactress – nobreak

933 – statrak karambit – phrase

**Training Dataset Analysis**

**14 no break hashtags**

**3 hashtags then a word**

**4 sets of consecutive hashtags broken up**

**20 phrases**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**VALIDATION DATASET**

9 – nonfiction stories collection – phrase

87 – roger smith hotel – phrase

94 – cfa risk mgmt – phrase

103 – lapierre btg – phrase

117 – bacon cheeseburger perogies – phrase (I think it’s one thing?)

118 – pumpkin perogies – phrase

142 – explicit – no break

148 - #iconlounge – no break

149 – trinity ave – phrase

162 - #amjadsabari – no break (person)

164 - #partywithpatti – no break

167 - #hananeelkhader – no break (person)

170 - #4thofjulysales – nobreak (event)

181 – t/y usher – phrase

189 - #teenchoice – no break

190 - #choicenextbigthing – no break

198 – prime minister – phrase

211 – fan up – phrase

225 - #brewers minor league – hashtag then 2 subsequent words into a phrase

232 – bloody mary - phrase

242 - #bernielovewave – nobreak

246 - #cherrybay party – hashtag then a word

277 – hdwill ep. – phrase

292 – it’s Korean

310 – drum island during ride – phrase

360 - #secondshooter – no break

369 – krusty krayayeah – phrase

374 – dp1756 ep ws9625 – phrase (episode numbers)

387 - #pokegobike – no break

423 - #supersoulsunday – no break

429 – lumen led – phrase

430 – xml t6 torch – phrase

435 – touch ups – phrase

**Validation Dataset Analysis**

**12 no break hashtags**

**2 hashtags then a word**

**0 sets of consecutive hashtags broken up**

**18 phrases**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**TESTING DATASET**

84 – taco shell – phrase

88 – amateur radio - phrase

106 – dentist procedures - phrase

140 – Wednesday special – phrase

145 – x8w x8c x8g rc quadcopter banana – phrase (But this is listing multiple products)

153 – morse code – phrase

157 - #marwat saeb – hashtag followed by a word

169 – scrubbing dish cloth – phrase

188 – t20 cricket – phrase

228 – one direction – phrase (not labeled properly)

231 - #barunsobti – no break (person name – not labeled properly)

232 - #sanayairani – no break (person name – not labeled properly)

241 - #momentsthatcount – no break

242 - #nissanpavilion – no break

250 - #kanhaiyakumar – no break (person name – labeled properly)

252 - #lettercarrier – no break

267 – early bird – phrase

278 – rajya sabha – phrase

318 - #rhino sanctuary – hashtag followed by a word

319 – eastern black rhino – phrase

325 – black wig brasil – phrase

363 - #thenextgalaxy – no break

391 – la provence – phrase

403 - #strictlybidness – no break

404/405 - #reverbnation #hitplay – consecutive hashtags that has been broken up

453 - #teenchoice – no break

454 - #choicemodel – no break

467 - #theassociation – no break

477/478/479/480 - #sentov #afanasyev #kolchencko #sososhenko – consecutive hashtags that has been broken up

494 - #tim peake – no break (person no detect)

499 - #gopdebate – no break

507 - #threatwire – no break

515 – meo cover house – phrase

554 - #shawnworldtourcologne – no break

555 – united states – phrase

561 - #july4canada – no break

565 - #teenchoice – no break

566 - #choicemodel – no break

596 – bishop father god – phrase

611 – sat July – phrase

621 – erectile dysfunction – phrase

628 - #southbelfast – no break (not labeled correctly)

650 - #purduecrew – no break

656 – eng sub – phrase

658 – nak tanya ni – phrase

661 – colouring page seahorse – phrase

663/664/665/666 - #sja #medicalfirstresponders #ontario #championship – consecutive hashtags that has been broken up

676 – oppo r9 cf – phrase

677 - #prettyp\*\*\*~ - no break

680 - #hellomynameis – no break

691 – mi abuela – phrase

**Testing Dataset Analysis**

**22 no break hashtags**

**2 hashtags then a word**

**3 sets of consecutive hashtags broken up**

**24 phrases**