Sporty Twitters

Virgile Landeiro Dos Reis

Machine Learning & Social Media – 11/21/2013

Problem Statement

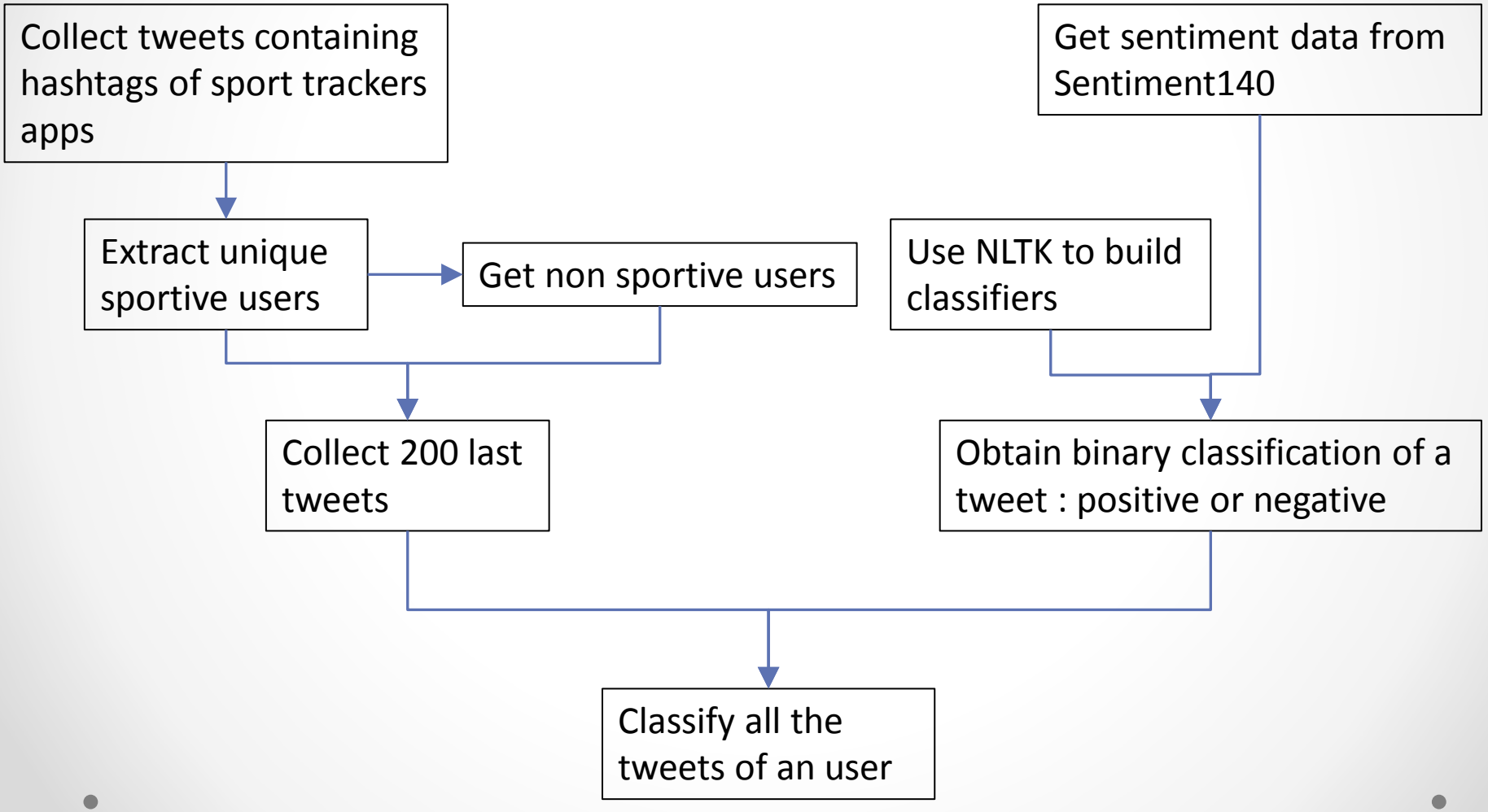
- Regular aerobic exercise can have benefits on health.
- Fewer studies focused on the psychological effects of exercising.

- Is it possible to correlate the physical activity of a person with his psychological well-being ?

Refined hypothesis

- The Twitter activity of those who actively use exercise trackers exhibits more positive indicators of psychological health.

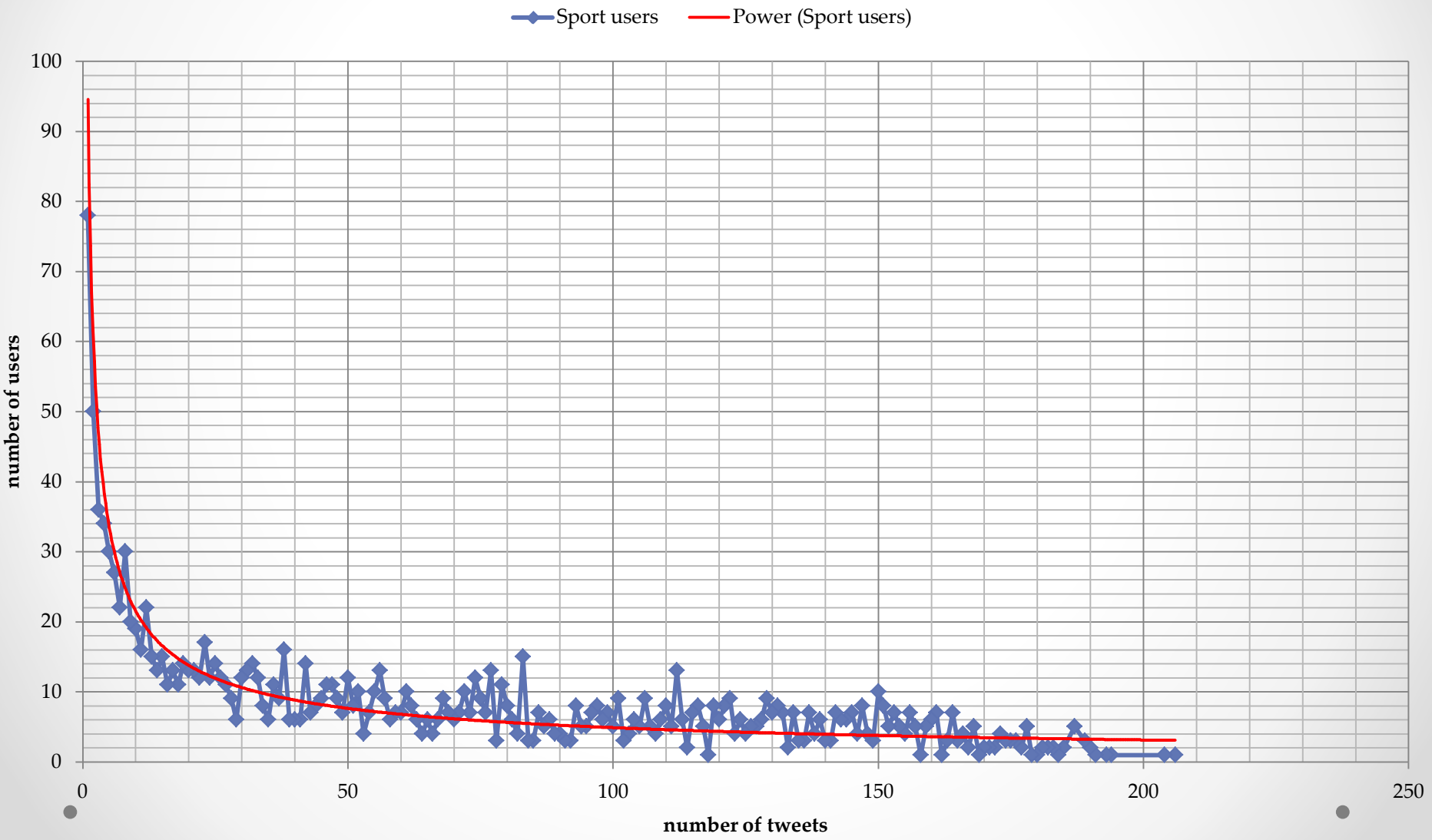
Method


Collect tweets containing hashtags of sport trackers apps
Get sentiment data from Sentiment140
Extract unique sportive users
Get non sportive users
Use NLTK to build classifiers
Collect 200 last tweets
Obtain binary classification of a tweet : positive or negative
Classify all the tweets of an user

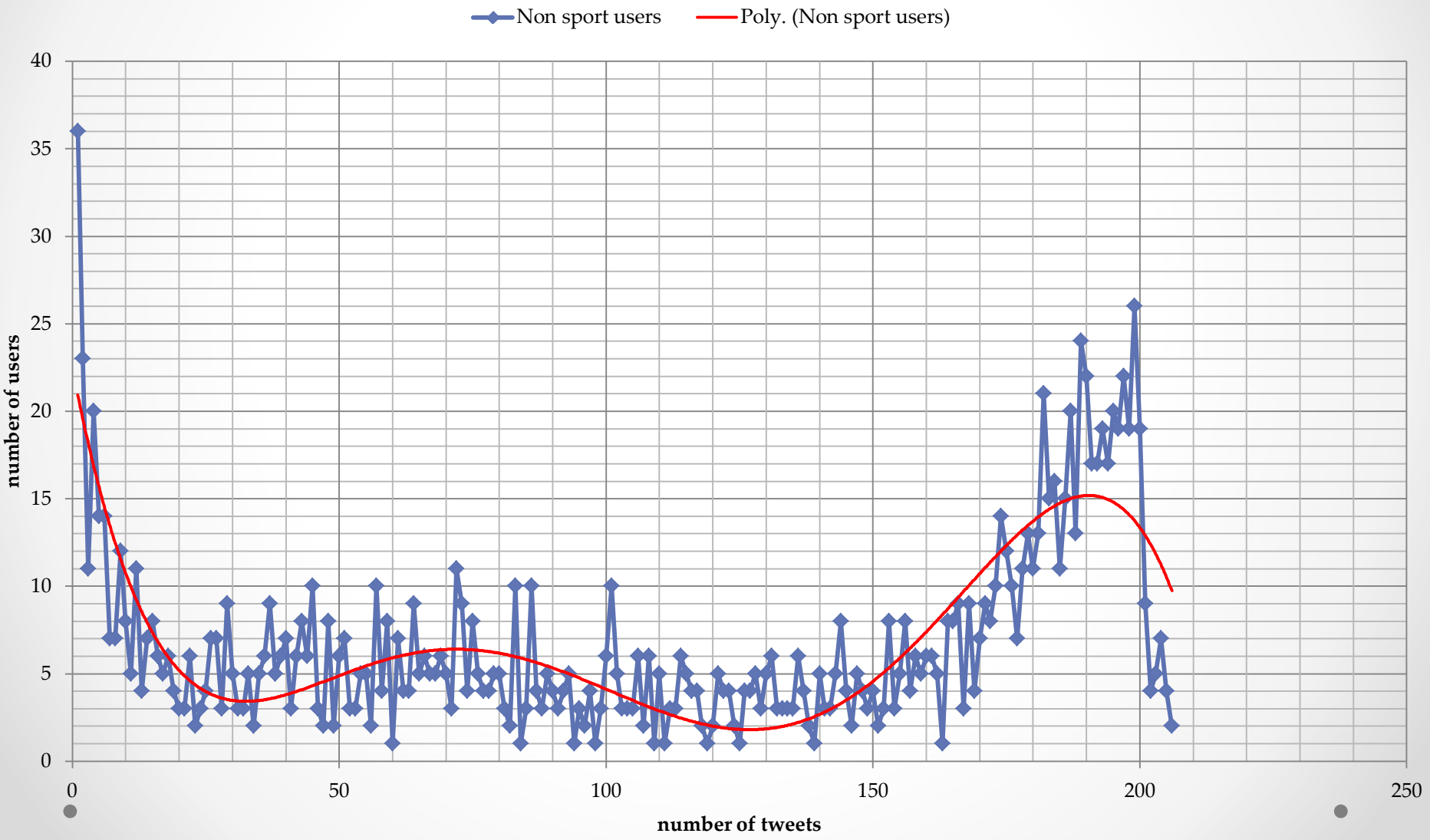
Data

- Collected 211,167 tweets by filtering sport trackers apps.
- Filtering: removed stopwords, removed users with more than 1,000 followers/followees, remove tweets generated by sport tracker apps
- Dataset after filtering:
  - 20,338 tweets by 1,579 sportive users
  - 28,070 tweets by 1,433 non sportive users

Sportive users distribution



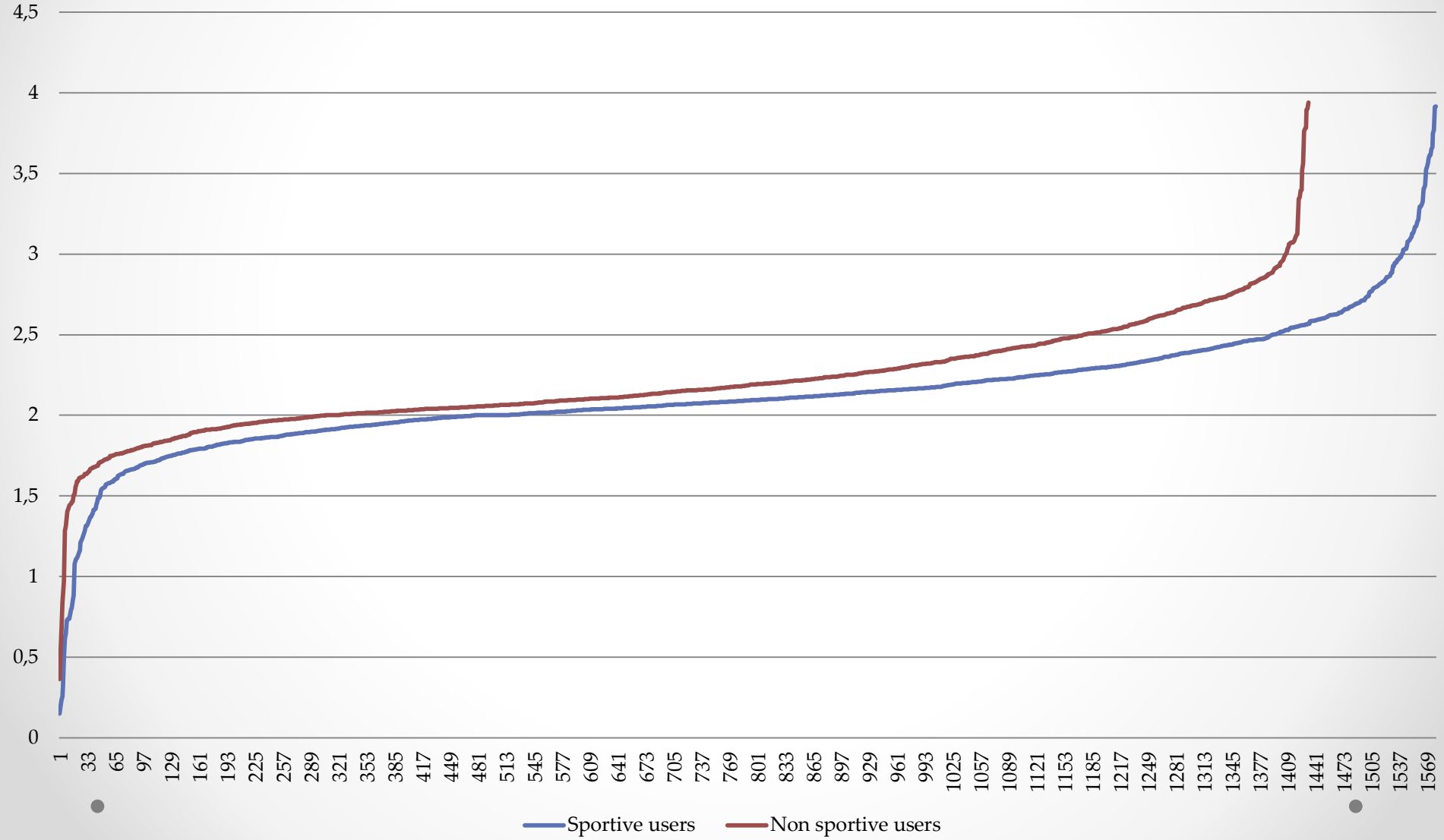
Non sportive users distribution



Classifiers

- Extracted all features (unigrams) from the Sentiment140 training data but training time too long
  ➔ use a cutoff by keeping only the N most frequent features of all the training data.
  ➔ Improved accuracy by avoiding overfit.
- Used Natural Language ToolKit in python to train several classifier over the features:
  - Naive Bayes classifier
  - MaxEnt classifier
- Created a MaxVote classifier using the three most « accurate » classifiers.

Results


4,5
4
3,5
3
2,5
2
1,5
1
0,5
0
1
33
65
97
129
161
193
225
257
289
321
353
385
417
449
481
513
545
577
609
641
673
705
737
769
801
833
865
897
929
961
993
1025
1057
1089
1121
1153
1185
1217
1249
1281
1313
1345
1377
1409
1441
1473
1505
1537
1569
Sportive users
Non sportive users

Future Work

- Collect new tweets and label them to create a well-being classifier (instead of a pos/neg sentiment classifier).
- Or use new labeled data to train the classifiers and get a better accuracy.
- Detect sportive users in an other way than just by collecting users that use sports tracker apps.

Thank you for your attention.

Questions ?