

TO WHAT DEGREE CAN WE USE NLP TO MINE CURRENT AND TRENDING TOPICS WITH RESPECT TO WELL-BEING?

Author
 Neel Manglani (n.manglani@student.tudelft.nl)
 Supervisors: Willem van der Maden, Garrett Allen, Ujwal Gadiraju & Derek Lomas.

Identifying what students engage with online could give us an indication of what students find relevant. Using technologies like NLP, it is possible to automate this retrieval of topics from popular online forums. Additionally, analyzing the sentiments of these opinions across core topics could give an indication of where the student-life zeitgeist trends towards overtime.

PROBLEM STATEMENT

More and more people online means an increase in user generated content. What if there was a way for someone to easily see which topics students engage with online? What if we could also see what the distribution of sentiment is for that topic?

This would be very beneficial for a group of survey makers that are interested in measuring well-being. For example, MyWellnessCheck[1] surveys various ranges of topics that students experience in their day-to-day lives. These include, but are not limited to, COVID-19, education, social issues, and physical health.

Furthermore, next to the core topics outlined in various models for well-being, it would also be beneficial to find 'trending' topics that are quite new in an automated way. (i.e: Will Smith Slap).

BACKGROUND

More and more digital activity

Finding the forums where students spend most of their time is key. A huge number of opinions is shared on a daily (and hourly!) basis. Mining these would provide us the possibility to understand the interests of students.

Word Embeddings

In order to create a notion of 'similarity, one can use word embedding. The spaCy package can convert a word into a 300 dimension vector[3], after which we can apply a distance function to signify 'similarity' between two words. Given a list of words, we can therefore group similar words into clusters using dimensionality reduction algorithms like PCA and so on.

Sentiment Analysis[4]

Next to mining WHAT students talk about, it would also be nice to know HOW they think about that particular topic. One year students might think negatively about their physical health, while another year could show an uptick in sentiment about that same topic.

RESULTS

	Negative	Positive
1	emotional breakdown	3,4,5th grade academic autism scholars
2	*war stop	workbooks
3	violation human rights	safe snacks
4	conspiracy theorist	surprise guests
5	blm activists	familiar faces heroes
6	critical race theory public education	support wellbeing
7	fyi america givetoomuch\$\$\$\$ ukrainewar	relaxation
8	sexual assault cover-up christian school	responsible business awards
9	misogyny	class school trips
10	slavery	creativity learning
11	craig calavetta	gap year programme
12	student loan crisis	informational interviews
13	domestic terrorism' letter biden admin	mentalhealthmatters
14	drug use	cyber security
15	no extra curriculars	special education workforce

figure 1: top 15 topics by intensity

	Negative	Positive
1	biden	education
2	america	high school
3	michigan ag	kids
4	police	parents
5	republicans	teachers
6	north carolina	money
7	summer school	high school baseball
8	school shootings	schools
9	republicans	student loans
10	hong kong	government
11	student debt	family
12	education activists	higher education
13	school teacher	summer
14	guns	business
15	kansas	grad school

figure 2: top 15 topics by frequency

METHOD

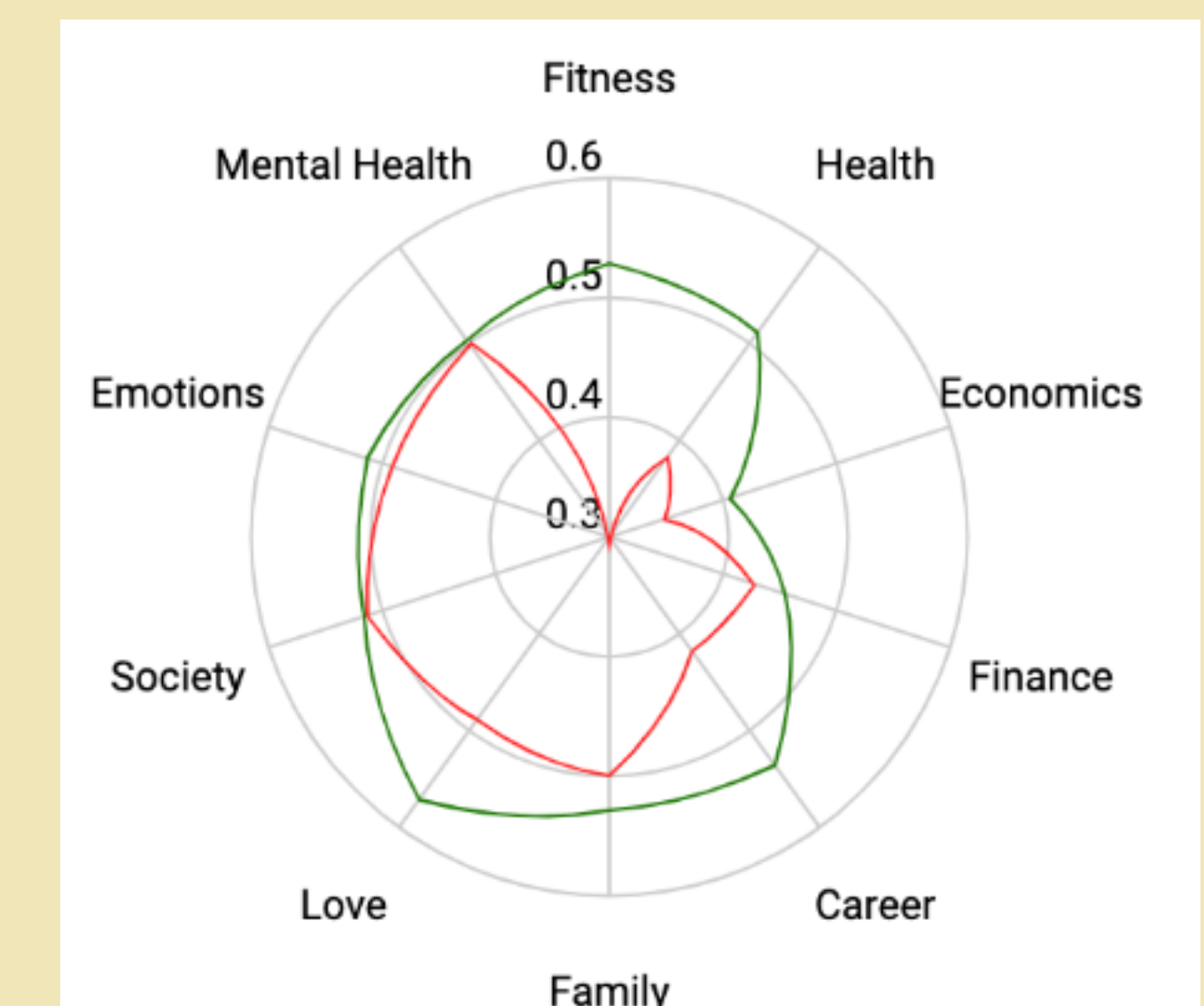
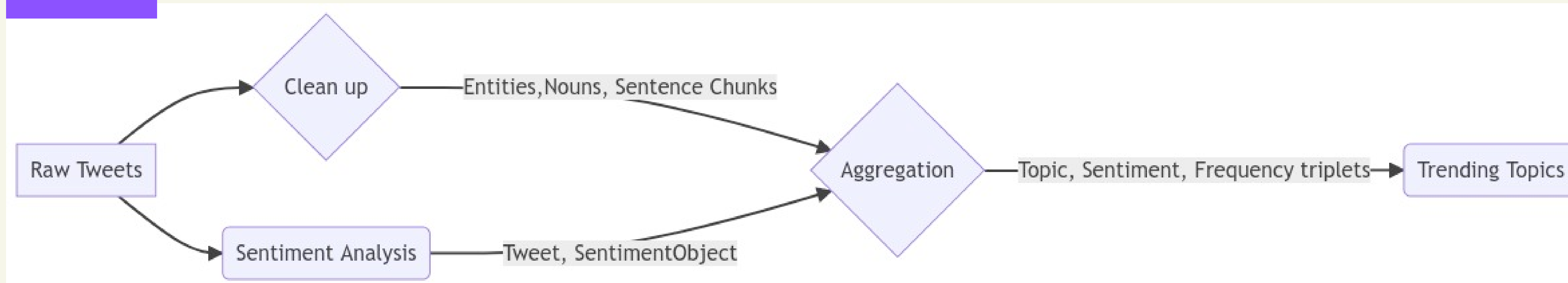


figure 3: well-being aspects against sentiment, green is positive, red is negative. 0 is low, 0.6 is max

Related literature

- <https://mywellnesscheck.org/>
- Jose Marquez and Emily Long. A global decline in adolescents' subjective well-being: a comparative study exploring patterns of change in the life satisfaction of 15- year-old students in 46 countries. Child indicators re- search, 14(3):1251–1292, 2021.
- https://huggingface.co/spacy/en_core_web_lg
- <https://github.com/cjhutto/vaderSentiment>

CONCLUSION

- 100'00 tweets were analyzed
- Biggest limitations were the input text, spam removal, and better preprocessing
- Interesting topics were discovered that might lead to questions that MWC[1] could use