To what degree can we use NLP to mine current and trending topics with respect to well-being?

Neel Manglani
Supervisor(s): Willem van der Maden, Garrett Allen, Ujwal Gadiraju, Derek Lomas
EEMCS, Delft University of Technology, The Netherlands

22-6-2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

The increase in global internet users brings forth a vast amount of social media users and therefore opinions that are shared online. A subset of those users, adolescents, seem to develop some sort of addiction towards social media, which could lead to low life satisfaction. This paper tries to extract trending topics and their relation to well-being in order to help organizations like MyWellnessCheck[1] check in on adolescents and students. The results indicated that this was possible, despite the vast amount of spam that is present online. Unsurprisingly, current events made the list of trending topics with negative sentiment like "school shootings", as well as unexpected topics with positive sentiment that could potentially improve well-being, like "safe snacks".

1 Introduction

As the number of internet users grow throughout time[2], it is expected that the sheer amount of online activity and therefore, user generated data, will increase alongside that[3].

However, life satisfaction of these users, especially young people, have been decreasing [4]. It has also been well established that a growing number of adolescents experience some sort of addiction to social media[5][6], and that this, argued by Hawi *et al.*, could lead to lower self-esteem and life satisfaction [7].

As these users often engage with content and topics that relate to interests that they have, it would be of great use to analyze these topics. Such a thing would be useful to have in order to identify topics that are *currently* very influential on these user's well-being. For example, organizations like MyWellnessCheck.org "want to make it easy for universities to check in" on students with respect to their well-being[1].

In the context of helping students keep track of their well-being, it is very important to keep up-to-date with the relevant topics that students talk about online. Think of areas like: passing examinations, maintaining an appropriate social life, or taking care of your health. Aside from this, it is also important to understand what the different subcultures of students are experiencing universally.

Fortunately, certain tool kits have been created in the recent past that can help us achieve these goals of analyzing social media posts. For example, word2vec[8] is a family of algorithms that can help us derive meaning from a given piece of text (i.e. a social media post). This will be thoroughly discussed in section 3.3. Additionally, by using algorithms like VADER Sentiment[9], one can classify a given piece of text with respect to positive, neutral and negative sentiment.

By combining (a) the vast amount of data that students provide in the form of posts on online forums with (b) algorithms that were described in the last paragraph, one can pose the question: To what degree can we use NLP to mine current and trending topics with respect to well-being?

This paper starts off by discussing a few related studies that have been conducted in other contexts in section 2. Their limitations are also discussed, as well as the gap that this paper aims to fill. Next, in section 3, a design for a pipeline is described that tries to solve the problem at hand. Section 4 show the results that have obtained by this pipeline, in which over 100‘000 tweets have been fed into. Lastly, section 5 and 6 discuss the results of this design and illustrate the limitations and possible future improvements that can be implemented.

2 State of Research

In the realm of Natural Language Processing and analysis of social media, the following great studies have taken place. Namely, Can et al.[10] illustrated how NLP was used to extract entities and the measurement of the sentiment of the stance towards those entities. Additionally, the tweets were mostly oriented around the Turkish language. This paper tries to apply this in the context of students and academia, which this study did not do.

Marshall et al.[11] showed how they used Twitter and NLP to analyze opinions about the pandemic using tweets issued by UK citizens. Whereas this study focused solely on British citizens and the impact of COVID-19 on their mental health, this paper will try to address the context of education. They also expressed the fact that their method of analysis was limited by the NLP platform that was chosen for their analysis, whereas the pipeline that this paper proposes in section 3.3 will be designed from the ground up.

Sendhilkumar et al.[12] also implemented sentiment analysis, but this time in the context of online products and books, and not online social media posts that were directly sourced from people in the world of academia, which this paper in turn will tackle.

This paper addresses the gap that has occurred via the factors that have been left out in the above-mentioned studies, and combines them, in order to reach an answer to the research question. In short, this paper will use NLP methods to analyze tweets in English made by student and faculty members of academic institutions all around the world. Additionally, topics will be extracted, in order to measure trends in the levels of sentiment found in these tweets.

2.1 Intended Contributions

One could imagine a world where survey makers like MyWellnessCheck.org [1] could simply log onto a dashboard and see what students are thinking about that day. A system to automate the retrieval of topics from posts made by people affiliated with education would prove extremely useful in order to "check-in" on a group of students. This would be a world where educational institutions and their faculty members could be in on the same page. For example, such a system would have ideally been able to pick up the trending

topic of ”working from home” early on during the COVID-19 pandemic, as that was quite the talking point online back then[11].

Aside from academia, this system could be used by large companies and their employees for the same purpose. Additionally, if applied properly, complaints and feedback from citizens could also be processed in order for government bodies to keep up with the various opinions.

In short, this paper shall use these new tools in order to create a newer tool that could help survey makers like MyWellnessCheck.org identify up-and-coming topics that are relevant to the lives of students.

3 Methodology

3.1 Data Collection

For the scope of this paper, Twitter was used as the primary source of data. The rationale being that Twitter is a form of open communication where text is the main format of information transfer, which comes in handy in order to extract topics from. This is in contrast to social media websites such as Instagram[1] which has a main focus on images, and TikTok[2] which emphasizes video content.

Therefore, by using the Twitter API[13], one had access to a vast amount of data in the form of text. Roughly 100‘000 tweets were fetched and stored. As the API provided a query-based system with which tweets could be looked up with, a search query was created. Marshall *et al.* stumbled across the same issue[11], where such a search query needed to be made to find relevant keywords for their goal of finding tweets related to the pandemic and mental health.

```
-has:media -is:reply -is:quote lang:en -is:retweet (student OR uni OR school OR college OR education OR #studentlife)
```

Figure 1: Query used as described in section 3.1 to fetch tweets related to education

Relevant keywords were accordingly identified, after which synonyms were created. Further changes to the query included the omission of ’retweets’, ’quotes’ and ’replies’ in order to capture genuine and personal opinions. Additionally, images, videos and other media were also omitted in order to focus on purely text-based tweets. Finally, the query included a clause where the language of the tweets had to be English. The keywords used for this tweet aggregation were school, student, uni, college and education. It was believed that these keywords would have encompassed a vast amount of tweets that related to education, life as a student, or educational institutions such as university, regular school, or college. For this paper, 100‘125 tweets (17MB of data) were ultimately used for analysis, which satisfied the query mentioned in figure 1.

[1] instagram.com

[2] http://tiktok.com

3.2 Data Preparation

The tweets were subject to a clean-up process. By removing stop words, lemmatization and applying further pre-processing, one could normalize the data in order to ensure some sort of consistency with respect to interpretation by a machine. For example, lemmatization, or stemming, refers to the conversion of a word into its original form that is specified in a dictionary. *Walking* resolves to *Walk* as *Cats* resolves to *Cat*. Stop word removal entails the removal of word that have no significant meaning in the broader context. These words include, but are not limited to: “a”, “are” , “the”, “is” and so forth.

Camacho-Collados *et al.* illustrated how useful this could be when working with domain specific corpora, as the corpus we are using exists in the domain of education. Their experiment compared not using lemmatization and using lemmatization. It was concluded that for domain specific corpora, the results obtained after only proceeding with the main processing were received ”poorly” [14] compared to including pre-processing. The next step was to choose a library to facilitate this. The system outlined in this paper eventually used the SpaCy[15] library. The effectiveness of this library was illustrated by Smelyakov *et al.*, as it was concluded that spaCy that showed a better output with respect to quality compared to other popular libraries after ”lemmatization (and removal of stop words)”[16].

After having applied the aforementioned clean-up procedures, the tweets were ready to be processed by the main pipeline. Figure 8 illustrates this with an example.

3.3 Data Pipeline

3.3.1 Sentiment Analysis

A key step in the pipeline was to extract sentiment. The goal was to ultimately have a measure for the degree to which a tweet expressed negative, positive or neutral sentiment, as well as a notion of *intensity* to express how strong this sentiment was.

Elbagir *et al.* argued that the use of the VADER[9] ”was an effective choice for sentiment analysis classification using Twitter data”[17]. To this extent, for the scope of this paper, the library of choice for sentiment analysis was VADER.

It is noteworthy to mention that this part of the pipeline was the only part that did not involve pre-processing of any kind. The reason behind this can be illustrated by considering a tweet with many exclamation marks, combined with the use of the word ”not”. As the exclamation marks would heavily contribute to the intensity of the tweet, while the word ”not”, a stop-word, could possibly indicate some sort of negation. Thus, removal of punctuation and stop-words would prevent indication of signal, as described in the documenta-

tion of VADER[3].

	Tweet	Sentiment	Intensity
1	My kid really goin to high school😭	negative	0.44
2	I GOT ACCEPTED INTO ART SCHOOL	positive	0.79
3	i like dogs	positive	0.1
4	I HATE COLLEGE	negative	0.99

Figure 2: Example of different sentiment outcomes

By doing the steps above, the tweets were tagged with not only values to indicate negative and positive sentiment, but also a *compound* value spanning from -1 to 1 to indicate a measure of *intensity*[18]. Refer to figure 2 to see the types of outcomes. For the scope of this paper, sentiment was determined using the compound value instead of the *neg*, *pos* and *neutral* values, which represented individual proportions. Note that throughout this paper, *intensity* refers to the absolute value of the compound value.

3.3.2 Entity Extraction

The next step in the pipeline was to extract relevant groups of words that have some sort of meaning in the real word. That is where Named Entity Recognition (NER) came in. NER is a sub-task of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, and so forth.

NER is performed by applying machine learning models to tokens in text that have been labeled with their part-of-speech and syntactic dependencies. See figure 3. The models are trained on a corpus of text that has been hand-annotated with named entities. After training, the models can be applied to new text to identify named entities.

Shelar *et al.* illustrated and concluded the following about the spaCy[15] library: "Python's Spacy gives a higher accuracy and the best result"[19] in terms of entity recognition. Aside from extracting entities, spaCy was able to provide access to a dependency tree for a processed corpus, similar to figure 3.

This feature was used to extract groups of nouns that formed a coherent entity. Combined with the aforementioned Named Entity Recognition process, the set of noun chunks and entities formed the set of topics that were extracted from a particular tweet.

[3]https://github.com/cjhutto/vaderSentiment

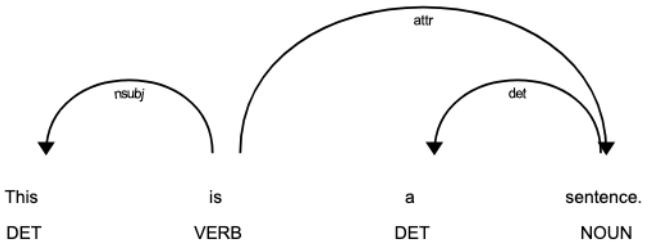

Figure 3: [Dependency Tree example]. Retrieved June 10, 2022, from https://spacy.io/usage/visualizers/#dep

3.3.3 Comparison to well-being

An important feature of NLP is of semantic similarity. This is where an algorithm like word2vec[8] came in. Word2vec is a neural network algorithm that is used to learn word embeddings from large amounts of text. The algorithm was able to take a corpus of text and produce a vector representation of each word in that given body of text. The vector representation then captured the context of the word in the text and could be used for various tasks such as text classification and similarity measurement. A classic example to illustrate contextual meaning would be as follows:

$$\vec{Amsterdam} - \vec{Netherlands} + \vec{India} \approx \vec{NewDelhi}$$

Similarity measuring being done, for example, by computing the cosine similarity (and therefore distance) between two vector representations.

Given the topics that were extracted from section 3.3.2, one could now start to rank them with respect to well-being. The Center for Disease Control (CDC) had fortunately compiled a list of concepts that can be interpreted as "the aspects of well-being"[20]. For the scope of this project, the following subset of aspects with were used to measure arbitrary concepts with respect to well-being:

	Aspect	Keywords
1	Physical	Fitness, Health
2	Economic	Economics, Finance, Career
3	Social	Society, Family, Love
4	Emotional	Emotions
5	Mental	Mental Health

Table 1: Core aspects of well-being as defined by the CDC and some keywords that describe them.

A few keywords were also associated with each aspect in order to help the Natural Language Processing algorithms. Now that these pre-defined aspects have been defined, one can illustrate the intended output. For example, consider the following body of text: "I cannot stress the importance of a

gym session before an exam". Our human intuition associates fitness and education with this text. The important question was, how could this process be automated? The general idea was to score the similarity between each topic in the *keyword* column of table 1 and that particular body of text. Ideally, the algorithm would have produced the highest score for aspect 1.

The equation to determine the similarity can be defined as follows:

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{ab}}{\|\mathbf{a}\|\|\mathbf{b}\|} = \frac{\sum_{i=1}^{n} \mathbf{a}_i \mathbf{b}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{a}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{b}_i)^2}} \quad (1)$$

In a nutshell, one compared the distance between two embeddings in an *n*-dimensional space.

The closer the embeddings are in vector space, the closer the semantic meaning is. Next, the core well-being keywords outlined in Table 1 were considered, as well as the tweets collected by the Twitter API outlined in section 3.1. Keeping in mind that the end goal was to gauge what these Twitter users were talking about, and the degree to which these conversations related to well-being, the following method was proposed ad hoc.

Initially, the largest model to contain pre-trained vectors was considered. This happened to be the spaCy[15] "en_core_web_lg" [4] model. This model was trained on news and blog articles, as well as comments from various social media networks. The model contained 343'000 unique vector embeddings which were used throughout this pipeline. What followed was the measurement of text with respect to the topics. There were two ways that this was done.

1. $\vec{Entity}$ compared to $\vec{Keyword}$
2. $\vec{Tweet}$ compared to $\vec{Keyword}$

This resulted in a collection of word and sentence vectors which one could have stored and use as many times as one wanted in the future, without the need of processing it again. Note that an embedding for a sentence could be computed by the *average* vector of the word vectors that describe the sentence. After having converted the well-being aspects into word embeddings, they were compared to the two types of embeddings mentioned above. For example, $\vec{Meditation}$ would be more similar to $\vec{MentalHealth}$, and would be relatively less similar to $\vec{Economics}$. Thus, being able to give insight into how a body of text relates to the well-being concepts outlined in table 1. Please refer to figure 9 to see a high-level overview of the pipeline.

3.4 Data Analysis

At this point in time, 100'125 tweets were stored and processed. In accordance with the pipeline from section 3.3, each tweet was tagged with, but not limited to, an intensity measure as outlined in section 3.3.1, extracted topics for which the process was outlined in section 3.3.2 as well as a similarity measure for each keyword as outlined in section 3.3.3. What followed was an aggregation of this data for each individual entity. This was done by iterating over each tweet, and then iterating over each topic extracted from section 3.3.2, and storing each topic as an entry with the same value for sentiment and similarity as it's parent tweet. Finally, for each unique topic, the average value for sentiment and similarity was computed, using the frequency of that particular entity. See figure 4.

	Topic	Sentiment	Intensity	Frequency
1	moscow	negative	0.43	6
2	cyber security	positive	0.79	4
3	netflix	neutral	0.02	10

Figure 4: Example of a list of extracted topics with sentiment, intensity, and frequency

In order to answer the research question, a measure of *trending* and *current* was needed. This could be interpreted as 'what users are (1) *mostly* (2) *engaging* with (3) *currently*'. Aside from the similarity of entities with the aspects outlined in table 1, three measures were introduced. Firstly, to address point 1, the entities were tagged with the frequency with which they appeared throughout the tweet collection. Secondly, point 2 was measured by the average intensity of sentiment value. And lastly, point 3 was a given, as all the 100'125 tweets collected were posted by their respective users on the same day, and scraped very close to the publishing date of this paper. Topics that only appeared once were filtered out, in order to combat spam.

Finally, once this aggregation of data was completed, it was then possible to query this newly created data. As the data set was quite large, it was decided to only initially consider the topics which fell under the 99.95^{th} percentile value for frequency and intensity each, in order to have a picture of the most trending, and the most *intensely* regarded topics. However, because such numerous topics were remaining after this filter process, the top 15 were identified for this paper. The full list can be found in the repository linked to this paper [5].

In this way, it was possible able to gauge the degree to which the extracted topics and the tweets themselves were interpreted as *intense* and *popular* combined with its relation to well-being. Thus, using these metrics, the question "To what degree can we use NLP to mine current and trending topics with respect to well-being? " was attempted to be answered.

[4] 'https://spacy.io/models/en#en_core_web_lg'

[5] https://github.com/Neel738/cse3000-data

Furthermore, each tweet was tagged with a keyword from table 1 which had the highest similarity with the body of the tweet. Next, the frequency of each keyword was determined, alongside the average positive and negative intensity.

4 Results

There were over 80‘000 topics extracted from the collection of tweets. After filtering the topics that appeared only once, there were still thousands of topics left. The following tables show the top 15 topics that have passed this filtering process. A distinction was made between topics ranked by intensity and frequency, each clustered by positive and negative sentiment.

The topics in these lists were subject to additional manual filtering, as a lot of spam tweets were encountered. This is elaborated in section 5.2. At first glance, the topics that are listed seem to heavily correlate to current events as of June 2022. The school shooting in Uvalde[21] has contributed quite a lot to the topics extracted, as shown in table 2 and 3. For example, ”school shootings” , ”guns” and ”craig calavetta” were topics that were frequently mentioned with negative sentiment. Additionally, the crisis in Ukraine[22] made its way to the list via the topics ”war stop” and most notably ”fyi america givetoomuch $$$$ ukrainewar”. Note that these topics were extracted from tweets that have undergone processing, and thus not directly representative of the raw tweet text.

As for positive sentiment, there were unsurprisingly quite a lot of references to education. Most notably, entries such as ”safe snacks”, ”class school trips” and ”mental health matters” seem to be trending topics that express a lot of potential in terms of further inquiry.

4.1 Top 15 topics ranked by Frequency

	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

Table 2: Topics ranked by Frequency, for both negative and positive sentiment

4.2 Top 15 topics ranked by Intensity

	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

Table 3: Topics ranked by Intensity, for both negative and positive sentiment

4.3 Relation to well-being

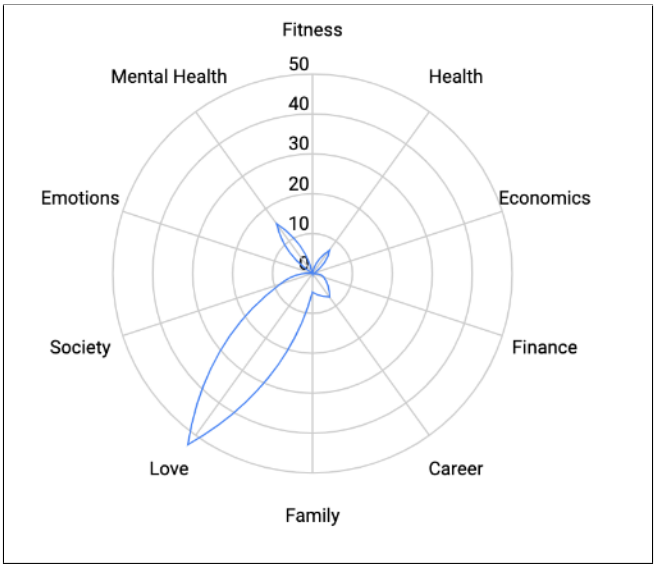

Figure 5: Keywords from table 1 against frequency, denoted in thousands

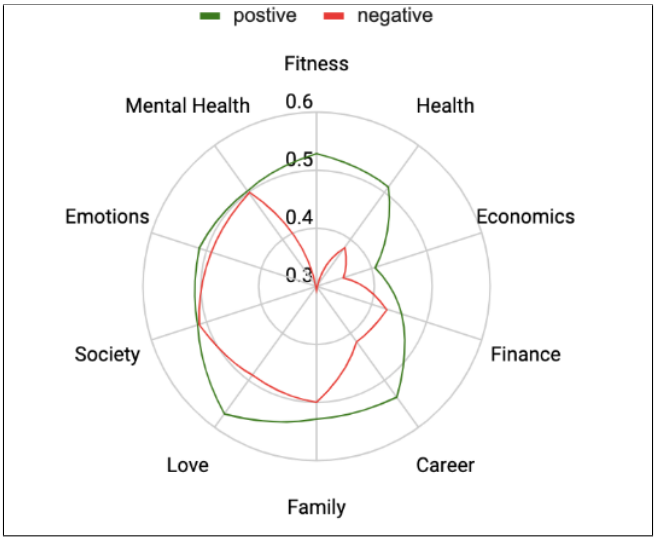

Figure 6: Keywords from table 1 against intensity of sentiment from 0-0.6, being the highest

When comparing the keywords that were used from table 1 with respect to frequency, it is apparent that by far, most of the tweets were most similar to the 'love' keyword. Which could indicate an interest in relationships. Mental health came in as second, followed by career and health. The rest of the keywords were negligible according to the interpretation in figure 5.

As for sentiment, there was a roughly equal distribution of positive sentiment amongst the keywords, except for economics, that scored relatively lower with respect to positive sentiment. On the other hand, figure 6 suggest that the scores for negative sentiment peaked in regard to mental health, society, love and family. Fitness, health and career were on the lower side of intensity, suggesting that these concepts were not as intensely regarded as *negative* compared to the others.

4.4 Relation to query keywords

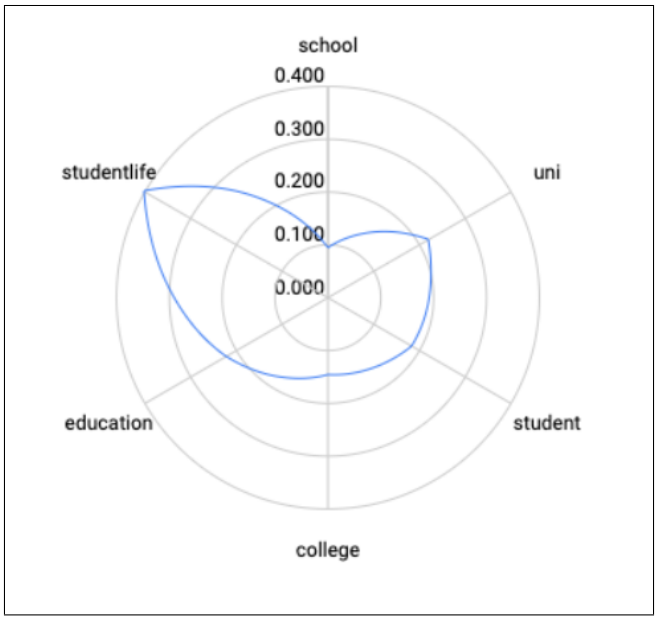

Figure 7: Filter keywords from 1 against intensity of sentiment from 0-0.4, being the highest

Some effort was done to analyze the intensity with respect to the query keywords used in figure 1. The results indicate a more intense overall sentiment with tweets containing the #studentlife keyword, whereas the school keyword scored the lowest.

5 Discussion

5.1 Findings

From the 100'125 tweets that were collected, processed and analyzed, about 83'000 topics were extracted. These topics ranged from dates, political parties, persons, geopolitical entities, to numbers. These topics were narrowed down to topics that were not only frequently mentioned, but also categorized according to sentiment.

The final list of topics that were outlined in section 4 provide an indication of what these Twitter users were talking about in a broader context. Current polarizing topics make the list such as the Uvalde shooting[21], the Ukrainian crisis [22] and opinions about the current sitting president. This was expected from the start, as current events form a part of every social media platform.

It is noteworthy to mention the topics that ranked high with respect to negative sentiment. Entries such as "student loan

crisis" and "drug use" could perhaps give one a peek into the mind of students and what currently bothers them the most.

What was not so expected was the sheer amount of spam accounts that have tweeted in the context of education. This took form by promoting various online courses, supplementary lectures, live streams, and so forth. A great deal of time was spent manually filtering through these types of tweets. What's more, is the appearance of similar topics in both the positive and the negative list. For example, "class school trips" and "no extra curriculars". The former having a highly positive score, while the latter scored a highly negative score. It would be worth investigating what exactly students would be looking for with respect to these seemingly ambiguous topics.

Finally, regarding the keywords outlined in table 1, one can interpret the data in figure 6 as follows: economics seems to be the least positively ranked keyword regarding sentiment, while mental health, emotions, society and family seem to be highly negatively ranked in terms of sentiment.

All things considered, the system was able to deliver a vast amount of topics that were current, trending as well as intense with respect to sentiment.

5.2 Limitations

There were many design choices considered for this system. In particular the decision to use Twitter as the input source. The fact that Twitter was the only data source used proves to be a limitation of the system. The tweets issued by the users do not represent the entirety of the student and academia members around the world. This can also be extended to other languages, in order to capture a broader audience and to provide an extra layer of dimensionality to the output.

Furthermore, as outlined before, the presence of spam and unwanted commercial tweets flooded the data set which made the collection of genuine opinionated tweets difficult to come by. Better pre-processing is therefore needed to combat this in the future. This could also entail the experimentation of different models that were trained using raw social media posts.

Another limitation of this system was the omission of retweets, likes and replies. These tend to be key indicators of engagement, that could potentially be incorporated in the ranking of the topics. These were left out due to time constraints.

Thus, there seemed to be many ways that this system proved to be limited. On the other hand, these limitations could become the basis for subsequent studies.

5.3 Responsible Research

Data obtained during the creation of this system was done so with the privacy of the user in mind. The tweets were anonymized by means of omitting the author's name from any storage and processing. All tweets that were processed are publicly available on the internet and via the Twitter API. Therefore, no immediate and first-hand contact was made with the publishers of the tweets, and no approval from the Human Research Ethics was sought after.

In terms of reproducibility, the system was designed in a way such that the implementation details were abstracted away. The major variables were the source of the data, the models that were used, along with the processing methods outlined in section 3.3. Great care was put into the documentation in order to assist anyone that would like to reproduce the system using any type of input.

6 Conclusions and Future Work

This paper tried to answer the question: To what degree can we use NLP to mine current and trending topics with respect to well-being? Over 100'000 tweets were analyzed to extract current and trending topics, including the ones that were ranked highly intense with respect to sentiment. Natural Language Processing algorithms were used in order to achieve this goal. Prior to this however, various clean-up processes were applied that helped increase consistency in order for the algorithms to process the large amount of text.

However, there are some limitations that could be the basis of further research. For example, the detection and omission of spam within the data set, as well as the use of multiple sources and perhaps languages in order to find topics that are trending in other parts of the world.

The findings indicate that this system is indeed capable of retrieving relevant topics. Topics ranging form negative subjects such as "emotional breakdowns" to positive recommendations like "safe snacks" and "school trips". Organisations like MyWellnessCheck.org[1] could therefore definitely use such a tool to monitor trending topics online, in order to understand their target audience better.

A Overview of cleanup process

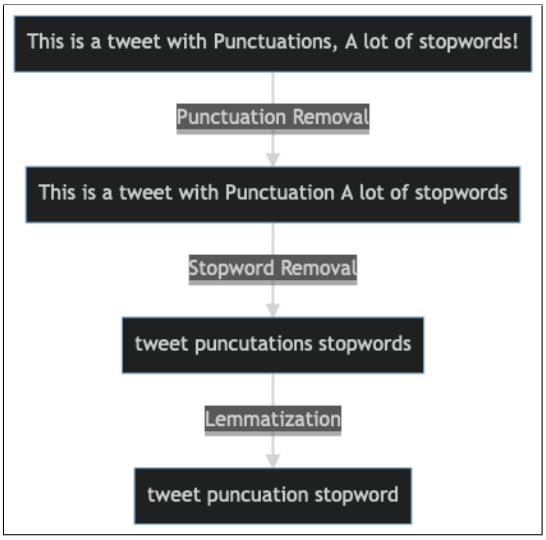

Figure 8: Example of cleanup process

B Overview of pipeline

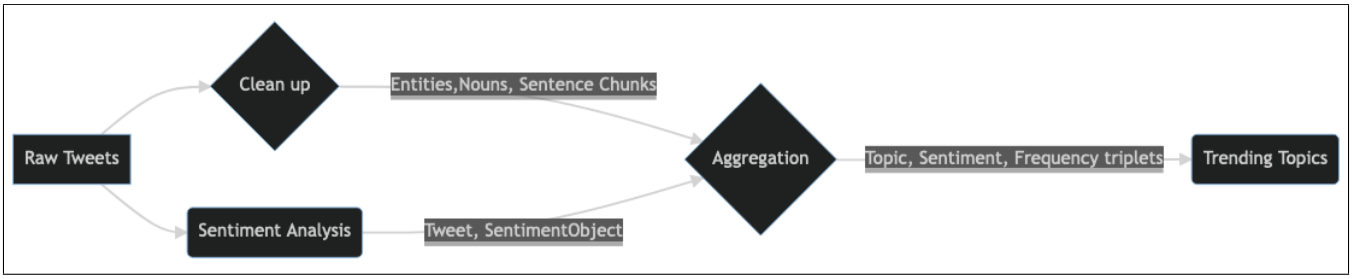

Figure 9: Overview of pipeline

References

[1] "Home new." https://mywellnesscheck.org/. (Accessed on 05/19/2022).

[2] "There will be 4.55 billion internet users worldwide this year — emarketer.com." https://www.emarketer.com/content/internet-use-worldwide. [Accessed 19-Jun-2022].

[3] "How Much Data Is Created Every Day in 2022? — techjury.net." https://techjury.net/blog/how-much-data-is-created-every-day/. [Accessed 19-Jun-2022].

[4] J. Marquez and E. 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 research*, vol. 14, no. 3, pp. 1251–1292, 2021.

[5] S. YAŞAR CAN and F. KAVAK BUDAK, "The relationship of social media use with depression and loneliness in adolescents: A descriptive study.," *Turkiye Klinikleri Journal of Nursing Sciences*, vol. 13, no. 4, 2021.

[6] M. D. Griffiths and D. Kuss, "Adolescent social media addiction (revisited)," *Education and Health*, vol. 35, no. 3, pp. 49–52, 2017.

[7] N. S. Hawi and M. Samaha, "The relations among social media addiction, self-esteem, and life satisfaction in university students," *Social Science Computer Review*, vol. 35, no. 5, pp. 576–586, 2017.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[9] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, pp. 216–225, 2014.

[10] D. Küçük and F. Can, "A tweet dataset annotated for named entity recognition and stance detection," *arXiv preprint arXiv:1901.04787*, 2019.

[11] C. Marshall, K. Lanyi, R. Green, G. C. Wilkins, F. Pearson, D. Craig, *et al.*, "Using natural language processing to explore mental health insights from uk tweets during the covid-19 pandemic: Infodemiology study," *JMIR Infodemiology*, vol. 2, no. 1, p. e32449, 2022.

[12] A. C. Arulselvi, S. Sendhilkumar, and S. Mahalakshmi, "Classification of tweets for sentiment and trend analysis," in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 566–573, 2017.

[13] "Documentation Home — developer.twitter.com." https://developer.twitter.com/en/docs. [Accessed 17-Jun-2022].

[14] J. Camacho-Collados and M. T. Pilehvar, "On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis," *arXiv preprint arXiv:1707.01780*, 2017.

[15] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing." To appear, 2017.

[16] K. Smelyakov, D. Karachevtsev, D. Kulemza, Y. Samoilenko, O. Patlan, and A. Chupryna, "Effectiveness of preprocessing algorithms for natural language processing applications," in *2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T)*, pp. 187–191, IEEE, 2020.

[17] S. Elbagir and J. Yang, "Twitter sentiment analysis using natural language toolkit and vader sentiment," in *Proceedings of the international multiconference of engineers and computer scientists*, vol. 122, p. 16, 2019.

[18] "GitHub - cjhutto/vaderSentiment: VADER Sentiment Analysis. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains. — github.com." https://github.com/cjhutto/vaderSentiment#about-the-scoring. [Accessed 18-Jun-2022].

[19] H. Shelar, G. Kaur, N. Heda, and P. Agrawal, "Named entity recognition approaches and their comparison for custom ner model," *Science & Technology Libraries*, vol. 39, no. 3, pp. 324–337, 2020.

[20] "Well-being concepts — hrqol — cdc." https://www.cdc.gov/hrqol/wellbeing.htm. (Accessed on 05/27/2022).

[21] W. F. Strong and L. Rodriguez Strong, "In memoriam: A tribute to the 21 lives lost in the uvalde school shooting (remembering the uvalde 21)," 2022.

[22] R. E. Mbah and D. F. Wasum, "Russian-ukraine 2022 war: A review of the economic impact of russian-ukraine crisis on the usa, uk, canada, and europe," *Advances in Social Sciences Research Journal*, vol. 9, no. 3, pp. 144–153, 2022.