MAL NLP Lexicon

Melancholy, Anxiety & Loneliness during lockdown

Tom De Smedt, Sofie Mariën, Guy De Pauw

MAY 2020

TEXTGAIN TECHNICAL REPORTS • TGTR2 • ISSN 2684-4842

AUTHORS

- **Tom De Smedt** (tom@textgain.com) has a PhD in Arts and is CTO at Textgain, focusing on Natural Language Processing. He was awarded the Belfius Smart Care Award for e-health in 2016.
- **Sofie Mariën** is researcher at the department of Communication Studies of the University of Antwerp. She specializes in social support among adolescents.
- **Guy De Pauw** has a PhD in Linguistics and is CEO of Textgain. He has over 20 years of experience in Natural Language Processing and Machine Learning.

ABSTRACT

We have created a new practice-based NLP resource for monitoring mental health on social media, in particular brooding. The resource is currently available for Dutch and captures 2,000+ expressions of anger, fear and sadness, along with various fine-grained mental states like despair, disappointment, hope, guilt, loneliness, melancholy, stress, relief and worry.

Keywords: *natural language processing*, *mental health*, *brooding*, *lockdown*, *coronavirus*

1 INTRODUCTION

Language is our projection of the self into society. We use nouns to name people and places, verbs to explain what we think and see, and adjectives to relate how we feel (De Smedt & Daelemans, 2012). What we say and how correlates with age and gender (Peersman, Daelemans & Van Vaerenbergh, 2011), personality (Verhoeven, Daelemans & De Smedt, 2013) and state-of-mind (Pennebaker, 2013). The Dutch MAL lexicon is a resource for Natural Language Processing (NLP) that captures brooding: feeling angry, down, sad or worried. For instance, people suffering from anxiety or depression will tend to use more absolute words (Al-Mosaiwi & Johnstone, 2018). The lexicon also has expressions of physical unease (e.g., *feeling tired*) that we use to avoid mental health taboo in Western societies.[1]

2 METHODS AND MATERIALS

The lexicon has 2,000+ Dutch words and word combinations that represent brooding. About 50% was collected from Twitter, using search queries with relevant verbs (e.g., *am*, *feel*), adverbs (*really*, *so*) and negative adjectives (*angry*, *bad*, *sad*, …), where we manually noted other relevant expressions in each message. About 25% was collected automatically from self-help discussion groups, using TF–IDF to identify keywords. In brief, TF is the number of times that a word occurs in a context (e.g., a self-help group) while DF is the number of times that it occurs in *any* context (e.g., Wikipedia, news articles). Dividing TF by DF reveals context-specific keywords. About 25% was collected by hand from student essays, where adolescents were asked to describe how they felt during the COVID-19 pandemic.

3 RESULTS

The lexicon groups words and word combinations into 8 different categories, each represented by an emoji and a four-letter alias that is easy to remember and tidy in programming code. There are 3 main categories for mental, physical and emotional expressions (MIND, BODY, SOUL). There are 3 emotional subcategories for anger, fear and sadness (RAGE, FEAR, LOSS), 1 mental subcategory for melancholy and depression (DOOM) and 1 physical subcategory for lethargy (DRAG).

MIND		Words that relate to **mental health** (e.g., *going insane*, *going nuts*), mental conditions (*panic attack, therapist*), negative thoughts (*brooding*) or rationalizing (*I see myself*).
BODY		Words that relate to **physical health** (e.g., *I feel bad*, *I feel weak*), physical conditions (*chronic pain*, *hospitalized*) or illness (*back pain, neck pain, stabbing headache*).
SOUL		Words that relate to **negative emotions** (e.g., *I feel angry*, *afraid*, *sad*), negative feelings (*grieving*, *totally worthless*) or craving (*really need*, *really want*).
RAGE		Words that relate to **feeling angry** (e.g., *gonna snap*, *the last straw*) or feeling frustrated (*I'm giving up, super annoying*, *super irritating*).
FEAR		Words that relate to **feeling afraid** (e.g., *panicking*, *panic*, *scares me*) or feeling anxious (*stress*, *this is a nightmare, this is a huge problem*).
LOSS		Words that relate to **feeling sad** (e.g., *heart breaks*, *heart weeps*, *passed*), feeling lonely (*miss you*, *need a hug*) or feeling alone (*no friends*, *not a sound*).
DOOM		Words that relate to **feeling down** (e.g., *dark days*, *dark thoughts, sigh*), feeling depressed (*dead inside*), feeling worthless (*I hate myself*) or feeling lost (*all my life*).
DRAG		Words that relate to **feeling tired** (e.g., *in bed all day*, *sleep all day*) or feeling bored (*watching the clock tick*).

[1] https://www.healthline.com/health/mental-health/physical-symptoms-of-depression

Each word or word combination in the lexicon also has an intensity score (0–4) and may be related to more than one category, as illustrated in the example below. The scores and categories were assigned manually by experts in linguistics and psychology.

SCORE	WORD(S)	🌧️	🤢	💔	🤯	😰	😭	😔	😳
★★★☆	*chaos in my head*	●	○	○	○	●	○	○	○
★★☆☆	*feel empty*	●	●	●	○	○	○	●	●
★★☆☆	*feel sad*	○	○	●	○	○	●	○	○
★☆☆☆	*irritating*	○	○	●	●	○	○	○	○

The score can be 0 (i.e., neutral but relevant), 1 (dramatic), 2 (dramatic and negative), 3 (fatalistic) or 4 (fatalistic to the point of alarming). We can use it to convert words to numbers, and then use statistics with those numbers. For example: “This irritating job makes me feel empty...” scores 1 + 2 = 3. A large collection of texts can be ranked by how alarming each text is.

This works even better if we count with exponential weights. The following weights essentially mean that a word with score 4 is as alarming as ten words with score 1 or a hundred words with score 0:

SCORE ▸	☆☆☆☆	★☆☆☆	★★☆☆	★★★☆	★★★★
weight	0.01	0.10	0.25	0.50	1.00

A demonstration of a concise Python script that imports the MAL lexicon and analyzes texts:

Listing 1. Example Python code for loading and using the MAL, using Grasp.py.

```
from grasp import csv, trie # https://github.com/textgain/grasp

mal = {}

for r in csv('mal.csv'): # (score, word, MIND, BODY, SOUL, RAGE, FEAR, LOSS, DOOM, DRAG)
    mal[r[1]] = {
        '0': 0.01,
        '1': 0.10,
        '2': 0.25,
        '3': 0.50,
        '4': 1.00,
    }[r[0]]

mal = trie(mal)

def rank(s):
    return sum(weight for i, j, word, weight in mal.search(s))

print(rank('niemand vindt mij leuk'))
```

4 ANALYSIS

In March and April 2020, during the ongoing COVID-19 pandemic many countries implemented travel restrictions and quarantines. Citizens were asked or forced to stay at home to slow the spread of the respiratory virus, and instead encouraged to use online social media to talk to family and friends. The effects of isolation and loneliness are well-studied and can contribute to depression in adolescents (Nangle, Erdley, Newman, Mason & Carpenter, 2003) as well as adults (Adams, Sanders & Auth, 2004).

Using the MAL lexicon, we were able to observe such an effect in the Belgian population. In the period from January 2019 to April 2020, we analysed 20,000 random Twitter messages each month written in Dutch (~300K in total). In general, higher scores are observed in November and January, also called the "darkest months" of the year. November has the highest number of RAGE expressions and January has the highest number of LOSS expressions. However, this does not hold for March and April 2020. At this time, messages should be scoring about as low as in summer, but this year they scored as high as in November or January, with 3x more BODY expressions (e.g., *cold*, *cough*, *fever*) and nearly 2x more matches in the categories LOSS, DOOM and DRAG, as the quarantine dragged on.

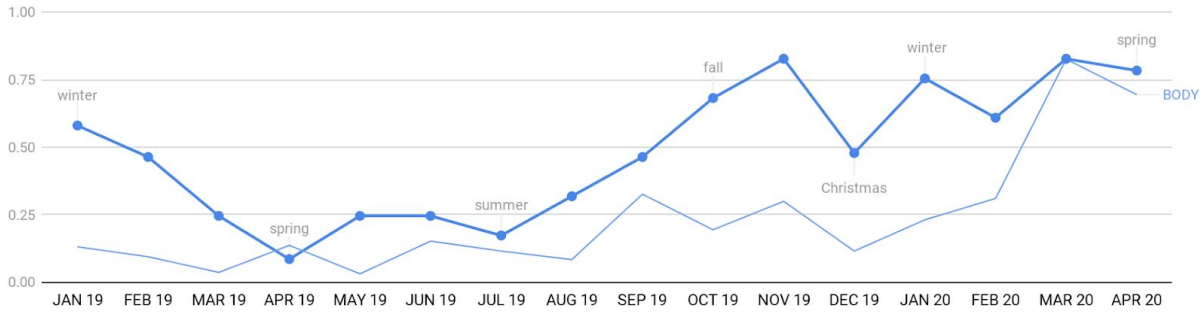

Figure 1. Negative thoughts & feelings (brooding) and physical complaints do not drop in spring 2020.

5 DISCUSSION

The Dutch MAL lexicon is a new NLP resource for verbal expressions of melancholy, anxiety and loneliness. It is not available for free. The reason is that there are now many examples of how such resources can be abused (cf. Facebook–Cambridge Analytica data scandal).[2] Academic research groups, societal NGOs and democratic governments can submit a request to get free access.

ACKNOWLEDGEMENTS

Co-funded by the Rights, Equality and Citizenship Programme of the European Union (REC).

REFERENCES

Adams, K. B., Sanders, S., & Auth, E. A. (2004). Loneliness and depression in independent living retirement communities: risk and resilience factors. *Aging & Mental Health, 8*(6), 475-485.

Al-Mosaiwi, M., & Johnstone, T. (2018). In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science, 6*(4), 529-542.

De Smedt, T., & Daelemans, W. (2012). "Vreselijk mooi!" (terribly beautiful): A subjectivity lexicon for Dutch adjectives. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC12), 3568-3572.

Nangle, D. W., Erdley, C. A., Newman, J. E., Mason, C. A., & Carpenter, E. M. (2003). Popularity, friendship quantity, and friendship quality: Interactive influences on children's loneliness and depression. *Journal of Clinical Child and Adolescent Psychology, 32*(4), 546-555.

Peersman, C., Daelemans, W., & Van Vaerenbergh, L. (2011). Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, 37-44.

Pennebaker, J. (2013). *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury.

Verhoeven, B., Daelemans, W., & De Smedt, T. (2013). Ensemble methods for personality recognition. In *Seventh International AAAI Conference on Weblogs and Social Media*.

[2] https://en.wikipedia.org/wiki/Facebook-Cambridge_Analytica_data_scandal