# Transgender Support Radar

## Using NLP to identify tweets in support of transgender rights

By Jessica Wulzen

# Our task

While LGBTQ+ rights have made tremendous progress, transgender acceptance remains contentious. Transgender people often live unsure if people around them are accepting.

We'll develop a tool that can scrape people's Twitter profiles to see if they have ever tweeted about trans-related topics, and if those tweets have been supportive or negative.

# Approach

1. Scrape all tweets from a given username
2. Identify which tweets, if any, are trans-related by using a list of terms
3. Perform sentiment analysis on trans-related tweets to classify negative/supportive
4. Report findings

The sentiment analysis is the main challenge, and thus is the focus of this presentation.

# Data Collection

I compiled 10,000 trans-related tweets and scored them from 1 (strongly anti-trans) to 5 (very supportive)

# Data Preprocessing

To classify tweets, we'll use typical text preprocessing conventions, with two exceptions

# Stopwords

Typically, stopwords (and, or, but, etc) are removed for NLP tasks

We need to keep some which are crucial in certain common phrases

**Trans women are women**

*After removing stopwords:*

**Trans women women**

**Trans women aren't women**

*After removing stopwords:*

**Trans women women**

By removing stopwords, these two messages with opposite meanings become indistinguishable from each other. Thus, some key stopwords will be kept.

Model improvement from this decision: 0.4% (every bit helps!)

# Quotation Marks

Typically, all punctuation, including quotation marks, is removed from text for NLP. However, in our context, they often show the speaker is not saying the words at face value:

*The word **“tranny”** is a slur - please don’t use it k thanks*

*My dad just said **“trans people are mentally ill freaks”**. He makes me so mad!*

*People saying **“trans women are women”** are delusional. Learn some biology.*

In each of these examples, the speaker uses quotes to show their disdain for the words in quotes. This should be captured in the model!

# Quotation Marks

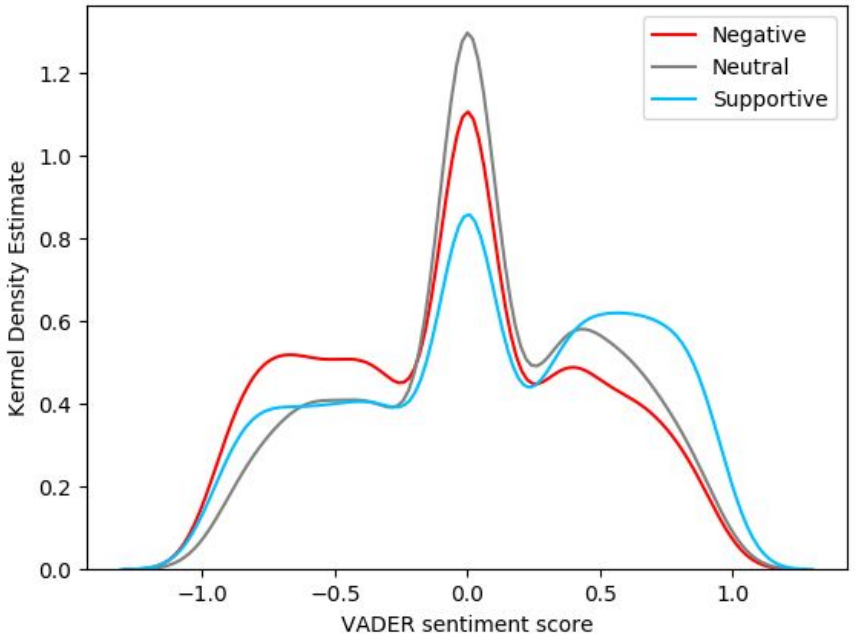
To handle quotation marks, we take each word in a quote and place them in their own quotes, so that quoted words are treated differently by the model:

*“trans women are women”*
becomes
*“trans” “women” “are” “women”*

Model improvement from this decision: 1% (quite significant!)

# VADER

VADER is a pre-trained model used for sentiment analysis, producing a score from -1 to 1 for the positivity of text - could it work for our task?



# VADER

Ultimately, VADER falls apart on ‘positive’ text with anti-trans messages, and ‘negative’ text with pro-trans messages:

*f*** transphobes! trans women ARE adult human females and if you disagree suck my **** : VADER of -0.92, but trans-supportive*

*Gender is binary. That’s the end of the topic. LGBT people please get help to overcome your twisted minds. God bless you all. : VADER of 0.84, but anti-trans*

A model using VADER gets 55% accuracy on classifying tweets, only 5% better than flipping a coin. Surely we can do better!

# TFIDF

To use the processed tweets in a model, we'll use a **Term Frequency-Inverse Document Frequency** vectorizer to turn the text into numbers

This approach weighs terms in a tweet relative to how many times it appears in a given tweet vs how commonly it appears in the entire dataset

We'll use phrases of length 1-4 words so that common multi-word phrases are properly captured

# Naive Bayes

We'll use Naive Bayes for classification, a model which works by asking the question for each term in a tweet: given that this term is in the tweet, what is the probability the tweet is supportive?

# Model Performance

On test data, our model correctly classifies 84.7% of tweets!

| | Predicted Supportive | Predicted Negative |
|---|---|---|
| Actual Supportive | 2185 | 358 |
| Actual Negative | 247 | 1173 |

# Most trans-supportive terms

| Phrase | Probability of supportive given tweet is just this phrase |
|---|---|
| tranniversary | 0.903 |
| nonbinary | 0.900 |
| trans men are men | 0.897 |
| genderqueer | 0.884 |
| trans rights | 0.874 |
| genderfluid | 0.862 |
| hrt | 0.847 |

The strongest predictors of a trans-supportive tweet are phrases typically only used by trans people, and nonbinary/trans male related terms (as trans-exclusionists usually focus on trans women)

# Least trans-supportive terms

| Phrase | Probability of supportive given tweet is just this phrase |
|---|---|
| woman adult | 0.053 |
| woman adult human | 0.054 |
| woman adult female human | 0.054 |
| natal | 0.075 |
| trannies | 0.084 |
| biological male | 0.085 |

The strongest predictors of an anti-trans tweet are variations on the phrase ‘adult human female’ used against trans women, slurs, and phrases which refer to trans women as ‘biological/natal males’

# Model issues

Language is complex, and many ways that people tweet are very difficult for our model to parse. Sarcasm is both very common and very difficult for Naive Bayes to detect. Quoting someone you disagree with tricks Naive Bayes because the meaning of the text is the opposite of the individual words.

*Oh sure, trans women are women! And trans men are men! And biology doesn't matter so we can ignore it and I can just decide to be a unicorn.*

*He literally said that trans people are dangerous predators and they should be rounded up and thrown in jail. He's just disgusting.*

# Full process in action

Let's see how the tool works by running the process on a real Twitter account.

The tool:

- Pulled all 641 tweets from the user
- Identified 9 that were trans-related (reading the remaining 632 confirms none were missed)
- Successfully classified 8 of those tweets
- Misclassified one tweet

A good maiden voyage!

# Next steps

As time permits, there are several further undertakings for this project, such as:

- More sophisticated modeling, using Keras with Word2Vec and neural networks
- Use the tool to gather data about tweeting trends among groups (US politicians, celebrities, etc)
- Host the tool online with a UI for easy use