

# Sentiment Analysis: It's Complicated!

Kian Kenyon-Dean<sup>1</sup>, Eisha Ahmed<sup>1</sup>, Scott Fujimoto<sup>2</sup>, Jeremy Georges-Filteau<sup>2</sup>, Christopher Glasz<sup>2</sup>, Barleen Kaur<sup>2</sup>, Auguste Lalande<sup>2</sup>, Shruti Bhanderi<sup>3</sup>, Robert Belfer<sup>3</sup>, Nirmal Kanagasabai<sup>3</sup>, Roman Sarrazingendron<sup>3</sup>, Rohit Verma<sup>3</sup>, Derek Ruths

1 - These authors contributed equally

2 - These authors contributed equally

3 - These authors contributed equally



**McGill University**

Primary contact:  
**Kian Kenyon-Dean**

*kian.kenyon-dean@mail.mcgill.ca*  
[https://github.com/  
networkdynamics/mcgill-tsa](https://github.com/networkdynamics/mcgill-tsa)



## Sentiment in Short-Text

*The entirety of **all human emotion** that could possibly be expressed in a piece of **short-text** can be captured by exactly **three categories**: Positive, Neutral/Objective, and Negative.*



# Sentiment in Short-Text

*The entirety of **all human emotion** that could possibly be expressed in a piece of **short-text** can be captured by exactly **three categories**: Positive, Neutral/Objective, and Negative.*

- *William Shakespeare (probably)*

# Sentiment in Short-Text

*The entirety of **all human emotion** that could possibly be expressed in a piece of **short-text** can be captured by exactly **three categories**: Positive, Neutral/Objective, and Negative.*

- *William Shakespeare (probably)*





# Verifying Shakespeare's Claim: New dataset

## MTSA: McGill Twitter Sentiment Analysis

- Collected 7,026 tweets: labelled with *5x coverage* (5 annotators per tweet)
- Selected from topics *Sports, Food, Media, Commercial Technology, General*
- Labelling crowdsourced with *CrowdFlower*



## **What sentiment is expressed in this tweet? Positive, Negative, or Objective?**

it took me 2 minutes to realize i was playing t  
pain



## What sentiment is expressed in this tweet? Positive, Negative, or Objective?

it took me 2 minutes to realize i was playing t  
pain

There is *no sentiment in this tweet* - it is neutral, or  
“objective”. 5/5 annotators labelled this as **objective**.



## **What sentiment is expressed in this tweet? Positive, Negative, or Objective?**

Who the hell is Lena Dunham don't tell me to  
Google I'm lazy.



## What sentiment is expressed in this tweet? Positive, Negative, or Objective?

Who the hell is Lena Dunham don't tell me to  
Google I'm lazy.

*Is there sentiment in this tweet?* 3/5 annotators labelled this as **objective**; 2/5 annotators labelled **negative**.



## **What sentiment is expressed in this tweet? Positive, Negative, or Objective?**

Oh my god you can control Chromecast via Google Home Mini which means Netflix without lifting a finger ever. Dangerous



## What sentiment is expressed in this tweet? Positive, Negative, or Objective?

Oh my god you can control Chromecast via Google Home Mini which means Netflix without lifting a finger ever. Dangerous

2/5 annotators labelled this as **positive**; 2/5 annotators labelled **negative**; 1/5 labelled **objective**.



## Labelling the Tweets: *standard approach*

| Tweet text                                                                                                             | + | - | O | Decision (?) |
|------------------------------------------------------------------------------------------------------------------------|---|---|---|--------------|
| I have the high ground. #StarWars                                                                                      | 0 | 0 | 5 | OBJECTIVE    |
| it took me 2 minutes to realize i was playing t pain                                                                   | 0 | 0 | 5 | OBJECTIVE    |
| Who the hell is Lena Dunham don't tell me to Google I'm lazy.                                                          | 0 | 2 | 3 | OBJECTIVE    |
| Cloud atlas was one interesting movie                                                                                  | 2 | 0 | 3 | OBJECTIVE    |
| Oh my god you can control Chromecast via Google Home Mini which means Netflix without lifting a finger ever. Dangerous | 2 | 1 | 2 | <i>Purge</i> |
| The irony of a vegan Thanksgiving dinner: No (animal) blood was shed in commemorating genocide in this meal            | 2 | 1 | 2 | <i>Purge</i> |



## Purging the MTSA (?)

- If we purged the data in the MTSA according to the standard of the OMD dataset (requires *majority* agreement), then **we would lose 7.9% of the data.**
- Requiring a minimum of *consensus* (4/5) agreement would result in **35.9% of the data purged.**
- If we purged the MTSA according to the standard of the STS-Gold (requiring *unanimous* agreement in order to “avoid noisy data”), **we would lose 65.5% of the data!**

| Agreement        | Count | % of Total |
|------------------|-------|------------|
| <i>Unanimous</i> | 2415  | 34.4       |
| <i>Consensus</i> | 2090  | 29.7       |
| <i>Majority</i>  | 1968  | 28.0       |
| <i>Disputed</i>  | 553   | 7.9        |
| <b>Total</b>     | 7026  | 100        |

Table 4: Annotator agreement rates. *Unanimous* stands for 100% annotator agreement, *Consensus* 80%, *Majority* 60%, and *Disputed* <60%.



## This seems a bit more... “Complicated”

### Purging the difficult data:

- In the classic Obama-McCain Debate dataset (Shamma et al. 2009), about 33% of the data gets purged due to annotator disagreement!
- In the STS-Gold dataset (Saif et al. 2013), about 26.5% of the data is purged!

*To purge, or not to purge, this tweet?*

Oh my god you can control Chromecast via Google Home Mini which means Netflix without lifting a finger ever. Dangerous



## Solution: add a “Complicated” label!

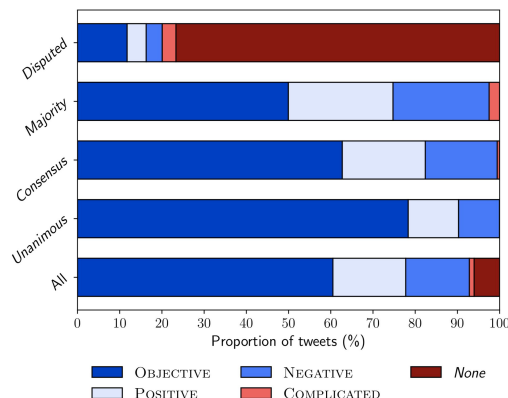
- Related to “Mixed” or “Other” in other datasets.
- Annotators were told:
  - Choose “Complicated” if the sentiment expressed in the tweet is ambiguous, mixed, or could be interpreted as both positive and negative.

### *5/5 Agreement on “Complicated”*

the iPhone 6s is so big  
and hard to use but I  
still like it

## Solution: add a “Complicated” label!

- Very few people thought things were “complicated”, individually.
- Only 0.9% of tweets had majority agreement on “Complicated”!
- Yet, 7.9% of the tweets had **no majority agreement** - *are these also “Complicated”?* Certainly they don’t seem “simple”!



| Label           | Count | % of Total |
|-----------------|-------|------------|
| OBJECTIVE       | 4186  | 59.6       |
| POSITIVE        | 1187  | 16.9       |
| NEGATIVE        | 1038  | 14.8       |
| COMPLICATED     | 62    | 0.9        |
| <i>Disputed</i> | 553   | 7.9        |
| Total           | 7026  | 100        |

Table 5: Distribution of tweets across classes, where the label given is the result of majority vote.



# Noisy annotators, or different data?

- *Experiment 1:* Set up task according to **standard practices**.
  - Purge **disputed** and “**Complicated**” data (8.75% of the dataset)
  - 3-class classification problem (*Positive vs Objective vs Negative*)
  - Use logistic regression classifier with standard sentiment analysis features. [Results from 5-fold cross-val.]



## Noisy annotators, or different data?

- *Experiment 1:* Set up task according to **standard practices**.
  - Purge **disputed** and “**Complicated**” data (8.75% of the dataset)
  - 3-class classification problem (*Positive vs Objective vs Negative*)
  - Use logistic regression classifier with standard sentiment analysis features. [Results from 5-fold cross-val.]
- **Twist:** test on the subsets with *different levels of annotator agreement*.
- **Hypothesis:** if results are the same across different levels of agreement, then the problem is bad annotators. Otherwise, **the data is qualitatively different**.



## Noisy annotators, or different data?

it took me 2 minutes to realize i  
was playing t pain (5/5 Objective)

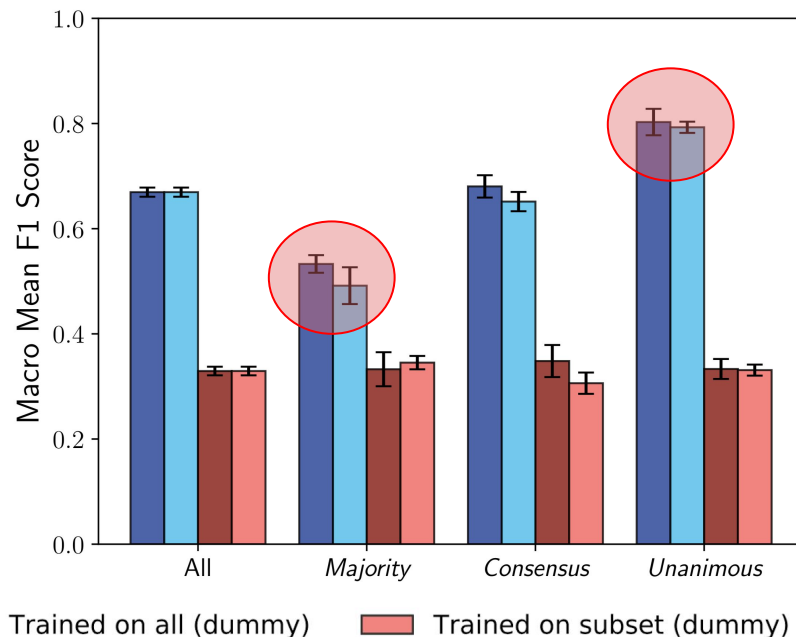
[?] == [?]

Who the hell is Lena Dunham don't  
tell me to Google I'm lazy. (3/5  
Objective)

- **Twist:** test on the subsets with *different levels of annotator agreement*.
- **Hypothesis:** if results are the same across different levels of agreement, then the problem is bad annotators. Otherwise, **the data is qualitatively different.**

# Noisy annotators, different data!

- Dramatically different results across agreement-level subsets!
- 53% accuracy on  $\frac{1}{3}$  (majority) annotator-agreement tweets;
- 68% accuracy on  $\frac{2}{3}$  agreement tweets;
- 80% accuracy on full agreement tweets!
- **Conclusion:** there is a **qualitative difference** across tweets of **different levels of annotator-agreement.**





## Can we detect “Complicated” data?

- Use the full dataset!
- Assign all “Disputed” (no majority agreement on label) as “Complicated”
- Assign all tweets with the label as their majority label, otherwise.

## Can we detect “Complicated” data? Not yet!

- Questions raised:
  - *Should we even assign labels according to **Majority** agreement?*
  - *What is the difference between “Disputed” and “Complicated”?*
  - *Would more advanced classifiers (RNNs, etc.) be able to detect this data better?*

| Label            | Precision | Recall | F1    |
|------------------|-----------|--------|-------|
| COMPLICATED      | 0.199     | 0.138  | 0.163 |
| POSITIVE         | 0.599     | 0.605  | 0.602 |
| OBJECTIVE        | 0.766     | 0.806  | 0.786 |
| NEGATIVE         | 0.510     | 0.487  | 0.498 |
| Total – weighted | 0.650     | 0.667  | 0.658 |
| Total – macro    | 0.518     | 0.509  | 0.512 |



**Should we care about “Complicated” data?**



## Should we care about “Complicated” data?

- Yes!



## Should we care about “Complicated” data?

- Yes!
- If anywhere between 8-30% of tweets are “Complicated” or have high levels of annotator disagreement, *should we really just throw them away?*
- If data with Majority vs. Unanimous agreement are qualitatively different, *should we not model and interpret them differently?*



# Should we care about “Complicated” data?

- Yes!
- If anywhere between 8-30% of tweets are “Complicated” or have high levels of annotator disagreement, *should we really just throw them away?*
- If data with Majority vs. Unanimous agreement are qualitatively different, *should we not model and interpret them differently?*

*Google & Apple would probably like to automatically interpret these!*

Oh my god you can control Chromecast via Google Home Mini which means Netflix without lifting a finger ever. Dangerous

the iPhone 6s is so big and hard to use but I still like it



## Conclusions and Perspectives

- Researchers should work with and release datasets with **the raw annotations**



## Conclusions and Perspectives

- Researchers should work with and release datasets with **the raw annotations**
- The raw annotations **may offer more informative signal for classifiers**



## Conclusions and Perspectives

- Researchers should work with and release datasets with **the raw annotations**
- The raw annotations **may offer more informative signal for classifiers**
- “Complicated” data should not be thrown away - it should be **understood!**



# Thank you!!

*Questions? Comments?*

kian.kenyon-dean@mail.mcgill.ca

Dataset and code: <https://github.com/networkdynamics/mcgill-tsa>