Centre for Research Training

Science Foundation Ireland For what's next

Sentiment Analysis

Message Polarity Classification

Task 4A from SemEval-2017: Sentiment Analysis in Twitter

Science Foundation Ireland Grant No- 18/CRT/6223

HOST INSTITUTION

UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh

PARTNER INSTITUTIONS

University of Limerick Ollscoil Luimnigh

NUI Galway Ollscoil na Gaillimhe

DCU

Trinity College Dublin Coláiste na Tríonóide, Baile Átha Cliath The University of Dublin

CRT·AI SFI Centre for Research Training in Artificial Intelligence

Team

Nivranshu Pasricha

Alex Randles

Victor Azevedo Coscrato

Chavvi Chandani

Mikhail Kudriavtsev

Overview

1. Task description
2. Related work
3. Data description and EDA
4. Challenges
5. Pre-processing
6. Implementation
7. Results

Task Description

Given a tweet, decide whether it expresses a POSITIVE, NEGATIVE or NEUTRAL sentiment

Sentiment Analysis

Sentiment

- Expression of emotions
- Mental attitudes or thoughts
- Social dimension
- Highly organized

Emotion

- Complex psychological states
- Psychological dimension
- Raw and natural

Levels

- Document
- Sentence (more fine-grained)
- Phrase/Word

Related Work

Pioneers

- **1999**: Subjective/objective sentences [Wiebe, Bruce, & O'Hara]
- **2002**: Classification of reviews using SVM and Naive Bayes [Pang, Turney]

Modern solutions

- **2017:** SVM [Liu], Naive Bayes, decision trees(ID3) [Phu]
- **2017**: CNN/RNN/LSTM with Attention [Baziotis, Pelekis & Doulkeridis]
- **2020**: Pre-trained models such as BERT [Devlin]

Other approaches

- **2013:** Association rule mining (lexicon of relationships between words) [Yuan]
- **2017:** Genetic algorithms (Lexicon optimisation) [Keshavarz]

Dataset Description

Twitter Dataset

Tweets are annotated for sentiment on a 3-point scale:

`POSITIVE`, `NEGATIVE` and `NEUTRAL`

637342059519680513	negative	Still bitter that they didn't tweet about MetL...
637691185100947456	positive	I remember buying my tickets to rowyso last ye...
637874723288936448	positive	@zourrysscheese hershey for tmh, on americas g...
638382158420307968	neutral	@uglyuni my mom already bought tickets and the...
6385339933448642 56	positive	Schreier Financial Services in Orange City wil...

Dataset containing tweet id, sentiment and text.

Twitter Dataset

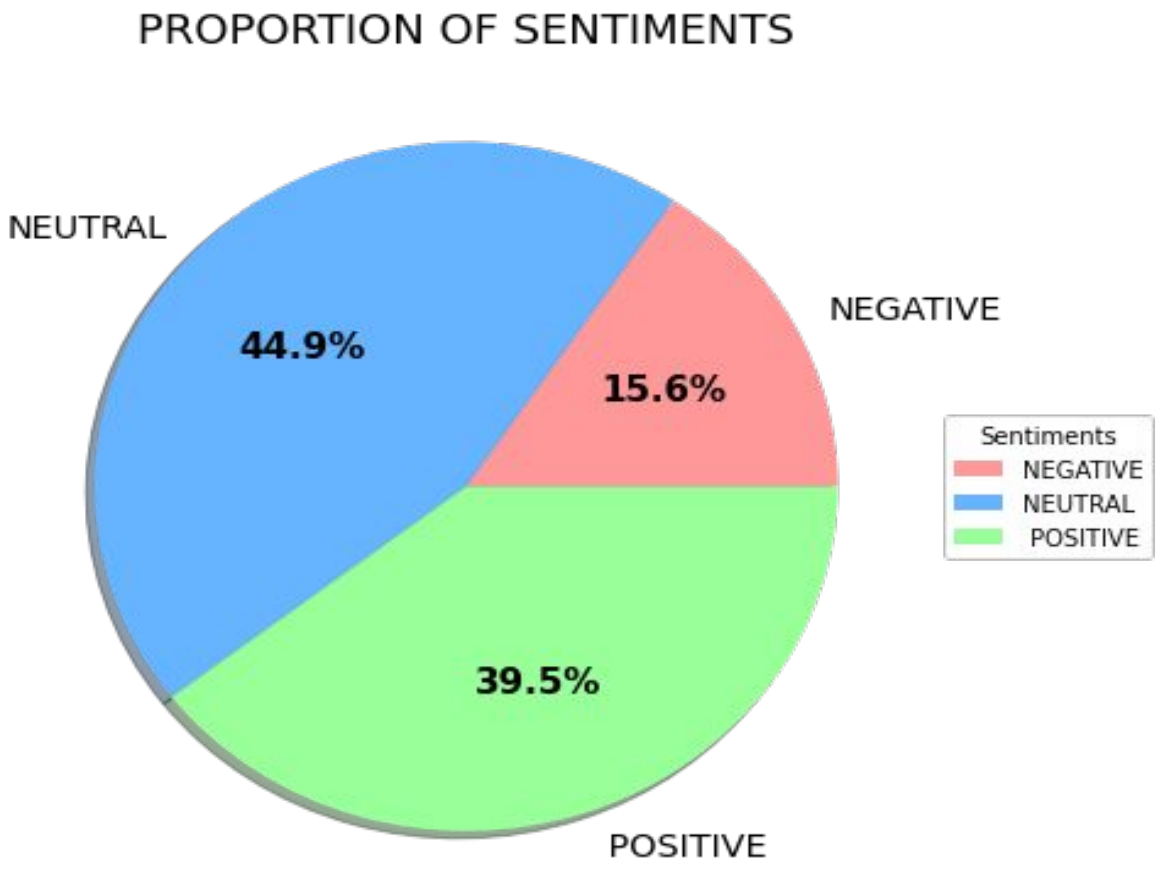
- Collection and Preparation:
  - Tweets from 2013-2017 [Nakov et al., 2013; Rosenthal et al., 2015; Nakov et al., 2016; Rosenthal et al., 2017]
  - Express sentiment about popular topics of the time
  - Techniques used:
    - Twitter-tuned NER system
    - Bag-of words
    - Manually filtering
    - SentiWordNet

Centre for Research Training

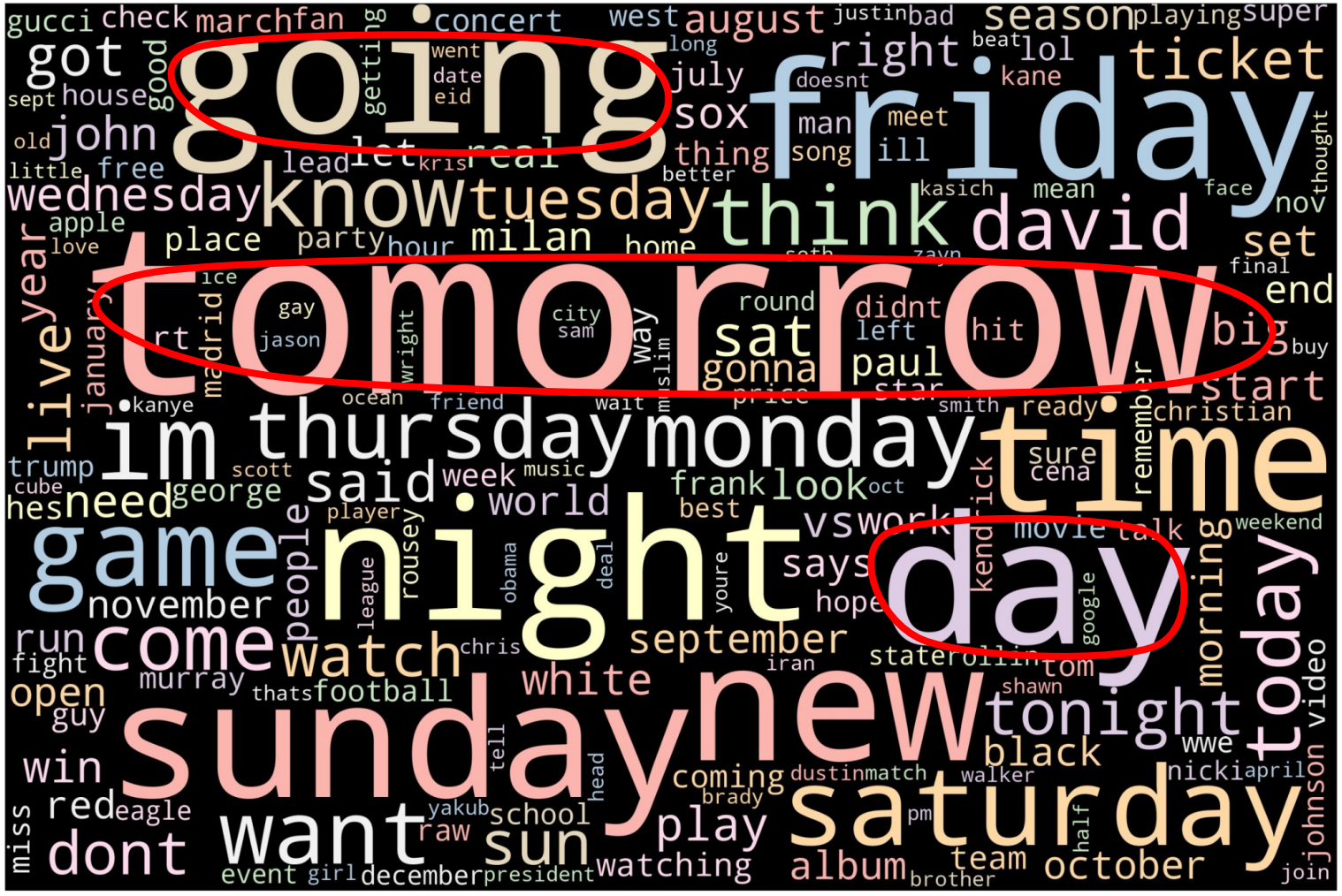
Science Foundation Ireland sfi For what's next

Exploratory Data Analysis

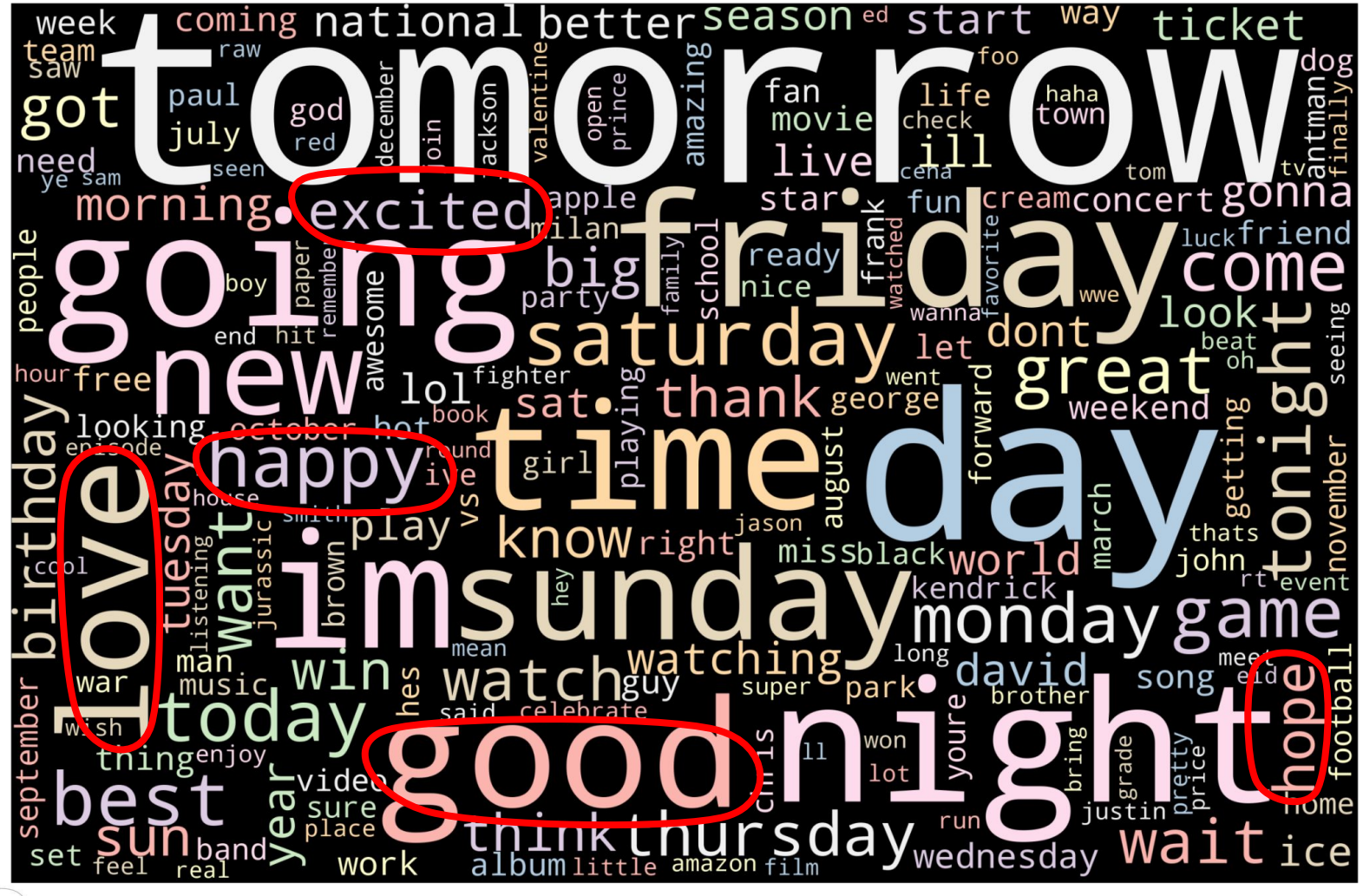
Sentiment Classes



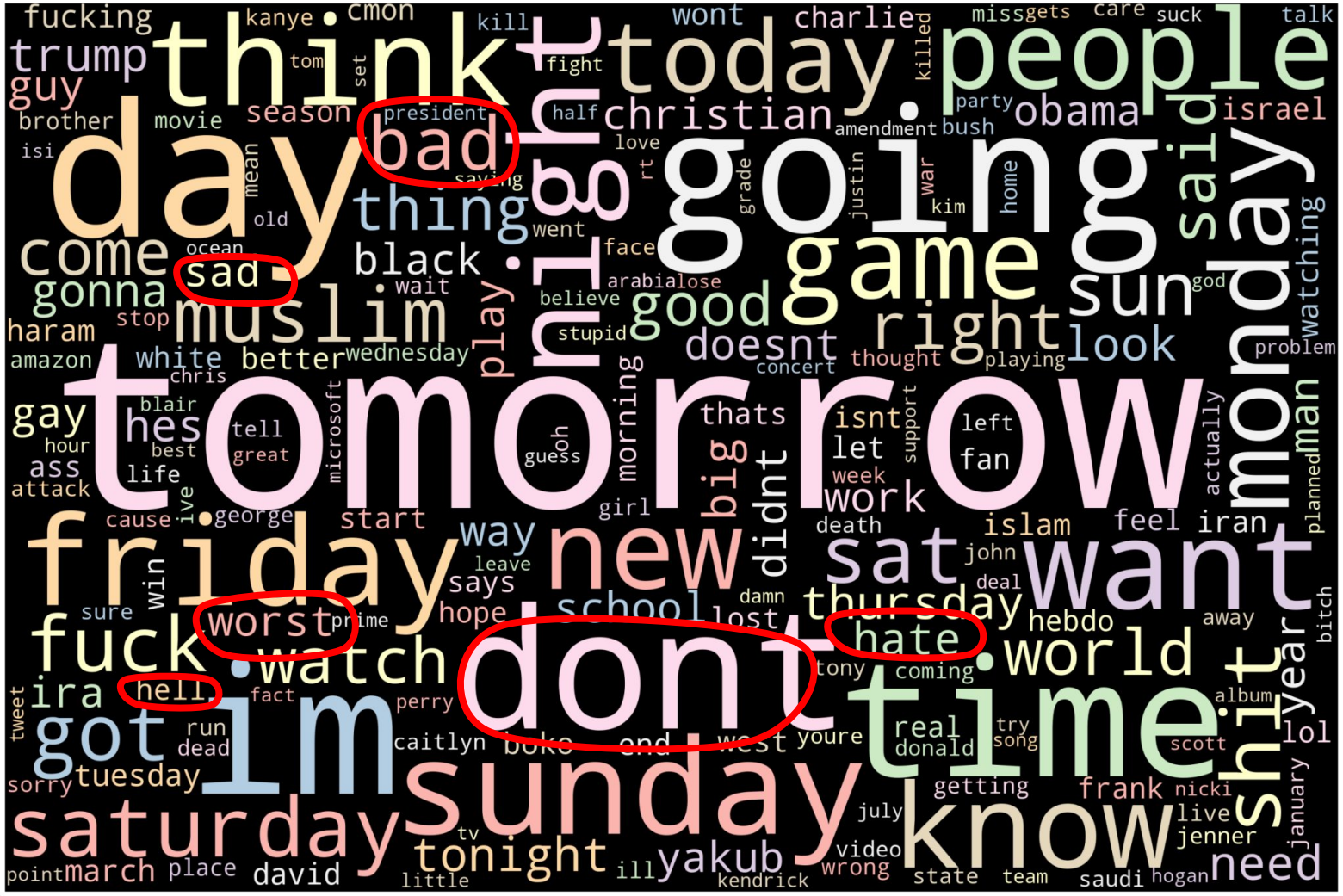
Neutral Tweets



Positive Tweets



Negative Tweets



Challenges in Tweet Classification

Misspelling

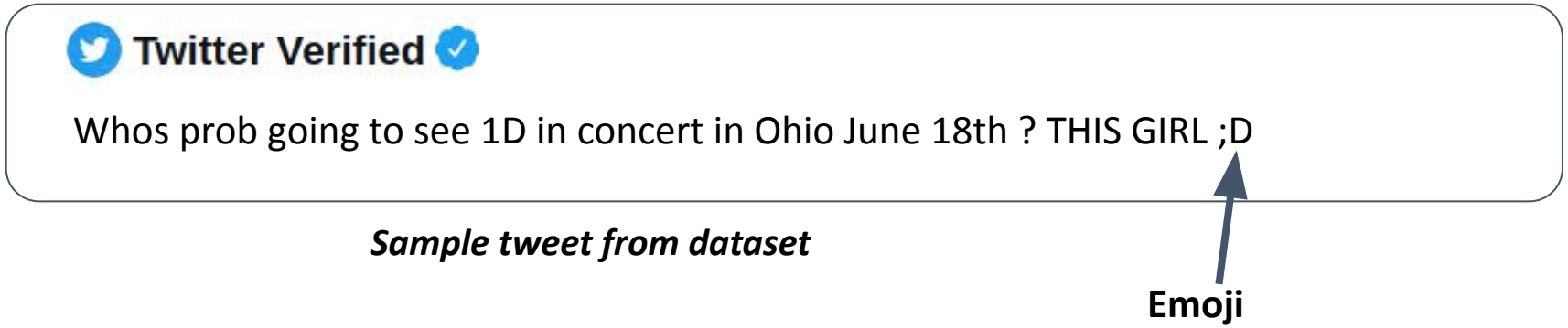
- Words found within tweets may not be in a lexicon

Sample tweet from dataset

Misspelled word

Emoji

- Emoji can provide further information about the tweet



Sample tweet from dataset

Slang

- Words and phrases that are regarded informal

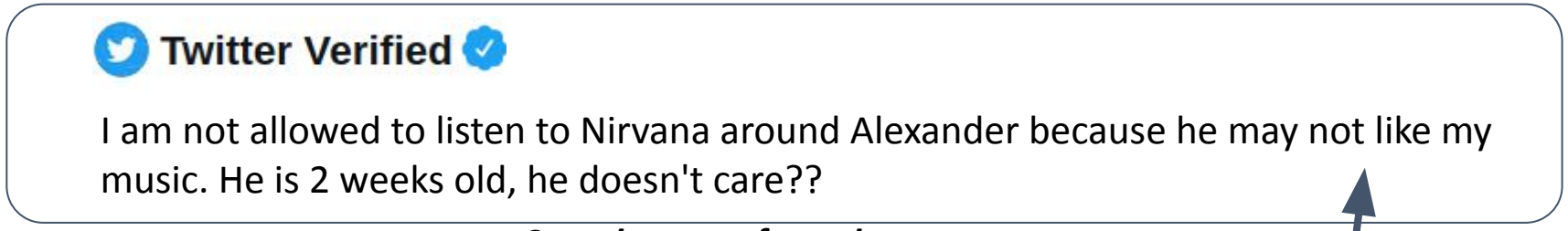
Sample tweet from dataset

Context

- Hashtags, replies and images often give more context about the tweet

Negation

- Positive words but negative sentence

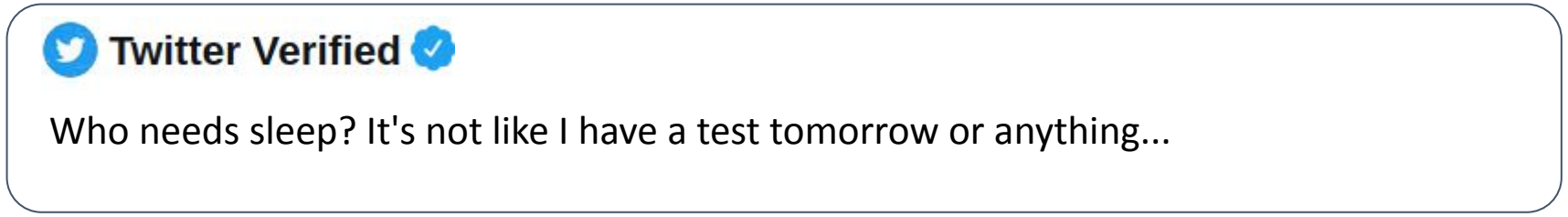


Sample tweet from dataset

Negation

Sarcasm

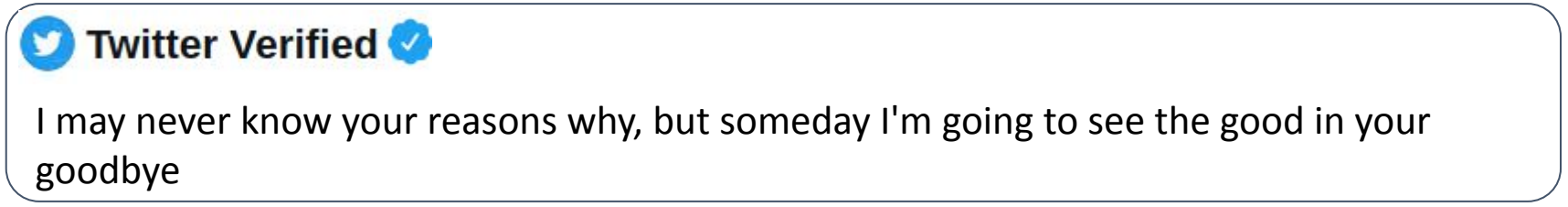
- Problem of determining if the actual meaning of a word is intended in a given tweet



Sample tweet from dataset

Implicit statement

- Neutral words used but implied statement



Sample tweet from dataset

Ambiguity

- Same words used with different sentiment

Sample tweets from dataset

Centre for Research Training

Science Foundation Ireland sfi For what's next

Implementation

Pre-processing Steps

- Misspelling:
  - Removing character repetition of length > *n* (`nltk.TweetTokenizer`)
- Emoji:

```
>>> emoji.demojize('Python is 👍')
'Python is :thumbs_up:'
```

- Slang:
  - Treat as it is!
- Hashtags:
  - Make sure to keep them

Techniques for Sentiment Analysis

- Rule-based Methods
  - Textblob - `PatternAnalyzer`
  - VADER: **V**alence **A**ware **D**ictionary for s**E**ntiment **R**easoning
- Probabilistic Methods
  - Naive Bayes Classifier using n-gram Features
- Deep Learning
  - BERT/BERTweet

Rule-based Methods

- Textblob - `PatternAnalyzer`
  - Uses a Subjectivity Lexicon for Adjectives

```
<word
    form="happy"
    cornetto_synset_id="d_a-9297"
    wordnet_id="a-01148283"
    pos="JJ"
    sense="enjoying or showing or marked by joy or pleasure"
    polarity="0.8"
    subjectivity="1.0"
    intensity="1.0"
    confidence="0.8"
/>
```

```
<word
    form="unhappy"
    cornetto_synset_id="n_a-521094"
    wordnet_id="a-01149494"
    pos="JJ"
    sense="experiencing or marked by or causing sadness or sorrow or discontent"
    polarity="-0.6"
    subjectivity="0.9"
    intensity="1.0"
    confidence="0.8"
/>
```

[https://github.com/clips/pattern/blob/master/pattern/text/en/en-sentiment.xml]

Rule-based Methods

- Textblob - `PatternAnalyzer`
  - Polarity (negative/positive, `-1.0` to `+1.0`)
  - Subjectivity (objective/subjective, `+0.0` to `+1.0`)
  - Words are tagged per sense:

```
<word
    form="ridiculous"
    wordnet_id="a-00752847"
    pos="JJ"
    sense="inspiring scornful pity"
    polarity="-1.0"
    subjectivity="1.0"
    intensity="1.0"
    confidence="0.9"
/>
```

```
<word
    form="ridiculous"
    wordnet_id="a-01266397"
    pos="JJ"
    sense="broadly or extravagantly humorous"
    polarity="1.0"
    subjectivity="1.0"
    intensity="1.0"
    confidence="0.9"
/>
```

Rule-based Methods

- VADER: **V**alence **A**ware **D**ictionary for s**E**ntiment **R**easoning [Hutton and Gilbert, 2014]
  - A Parsimonious Rule-based Model for Sentiment Analysis
  - Handle contractions and negations (e.g., “wasn’t very good”)
  - Punctuation to signal increased intensity (e.g., “Good!!!”)
  - Use of word-shape to signal emphasis (e.g., ALL CAPS)
  - Degree modifiers (e.g., “very”, “kind of”)
  - Slang words (e.g., “kinda”, “uber”)
  - Emoticons and UTF-8 emoji (e.g., 💘 and 😁)
  - Acronyms (e.g., “lol”)

Probabilistic Methods

- Naive Bayes Classifier
- Bayes Rule to estimate the probability of a label

$$P(\text{label}|\text{features}) = \frac{P(\text{label}) * P(\text{features}|\text{label})}{P(\text{features})}$$

- n-grams as features
- Assumption: Features are independent

[https://www.nltk.org/_modules/nltk/classify/naivebayes.html]

Most Important n-gram Features

Deep Learning

- BERT: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers [Devlin et al., 2019]
- BERTweet [Nguyen et al., 2020]
  - Contextualised embeddings for tweets
  - Trained using RoBERTa procedure [Liu et al., 2019]
  - 845M Tweets; 80GB of uncompressed texts
  - 5M tweets related to COVID19 (January to March 2020)
- Tokenization
  - https://crt-ai.ie → `<HTTPURL>`
  - @somebody → `@USER`
  - 😡 → “angry”

Evaluation

- 3 Official Metrics
  - Average Recall
  - Macro-average F1
  - Accuracy

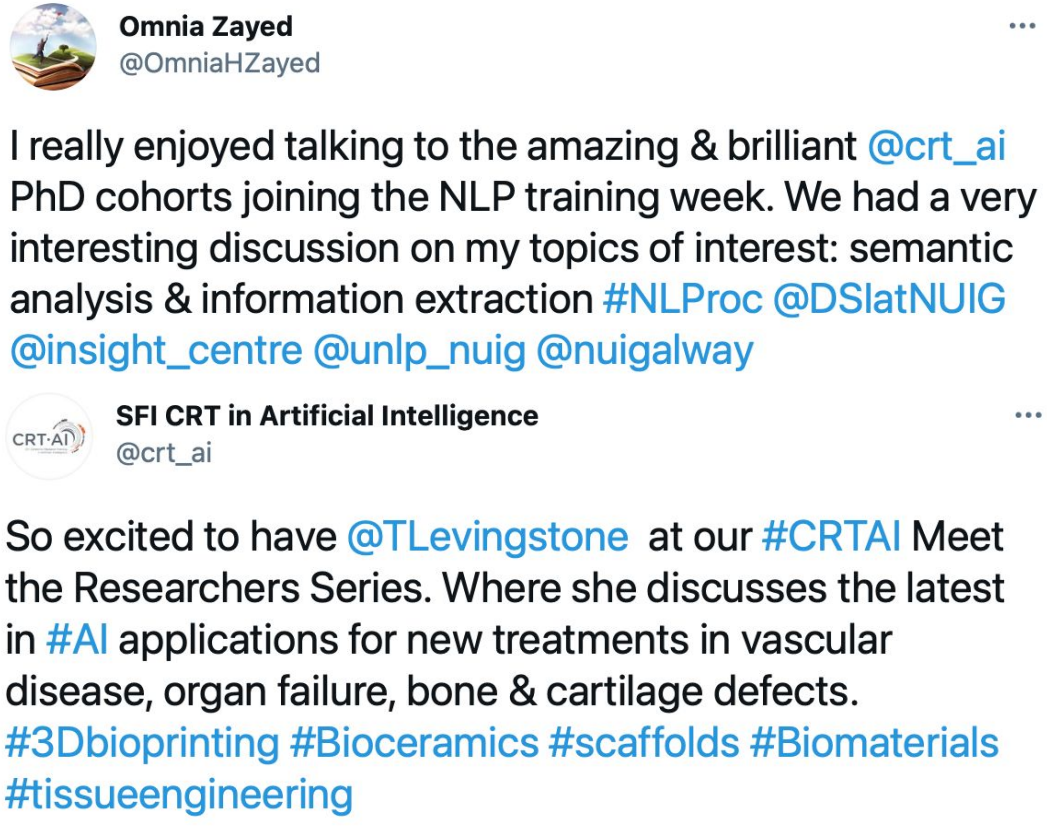
$$AvgRec = \frac{1}{3}(R^P + R^N + R^U) \qquad F_1^{PN} = \frac{1}{2}(F_1^P + F_1^N)$$

Results

System	Average Recall	Macro-average F1	Accuracy
LSTM+CNN ensemble†	0.681	0.677	0.651
Textblob	0.490	0.414	0.487
VADER	0.570	0.528	0.530
Naive Bayes	0.562	0.554	0.537
BERTweet	0.724	0.720	0.708

† Best performing system at SemEval 2017

Examples



	VADER	Textblob	Naive Bayes	BERT
I really enjoyed talking to the amazing & brilliant @crt_ai PhD cohorts joining the NLP training week. We had a very interesting discussion on my topics of interest: semantic analysis & information extraction #NLProc @DSIatNUIG @insight_centre @unlp_nuig @nuigalway	✅	✅	✅	✅
So excited to have @TLevingstone at our #CRTAI Meet the Researchers Series. Where she discusses the latest in #AI applications for new treatments in vascular disease, organ failure, bone & cartilage defects. #3Dbioprinting #Bioceramics #scaffolds #Biomaterials #tissueengineering	❌	✅	✅	✅

✅ Correct classification
❌ Incorrect classification

Examples

	VADER	Textblob	Naive Bayes	BERT
 **Omnia Zayed** @OmniaHZayed ··· I really enjoyed talking to the amazing & brilliant @crt_ai PhD cohorts joining the NLP training week. We had a very interesting discussion on my topics of interest: semantic analysis & information extraction #NLProc @DSIatNUIG @insight_centre @unlp_nuig @nuigalway	✅	✅	✅	✅
SFI CRT in Artificial Intelligence @crt_ai ··· So excited to have @TLevingstone at our #CRTAI Meet the Researchers Series. Where she discusses the latest in #AI applications for new treatments in vascular disease, organ failure, bone & cartilage defects. #3Dbioprinting #Bioceramics #scaffolds #Biomaterials #tissueengineering	❌	✅	✅	✅

✅ Correct classification
❌ Incorrect classification

Examples

	VADER	Textblob	Naive Bayes	BERT
Sarcasm	❌	❌	✅	❌
Negation	✅	✅	❌	✅

✅ Correct classification
❌ Incorrect classification

Conclusion

- Approaches for Sentiment Analysis
  - Rule-based and Lexicon-based Methods
  - Probabilistic Methods
  - Deep Learning
- More sophisticated models proved to be more accurate
- Always check your domain

Thank you for your attention

Code : https://github.com/pasricha/crt-nlp-week

39