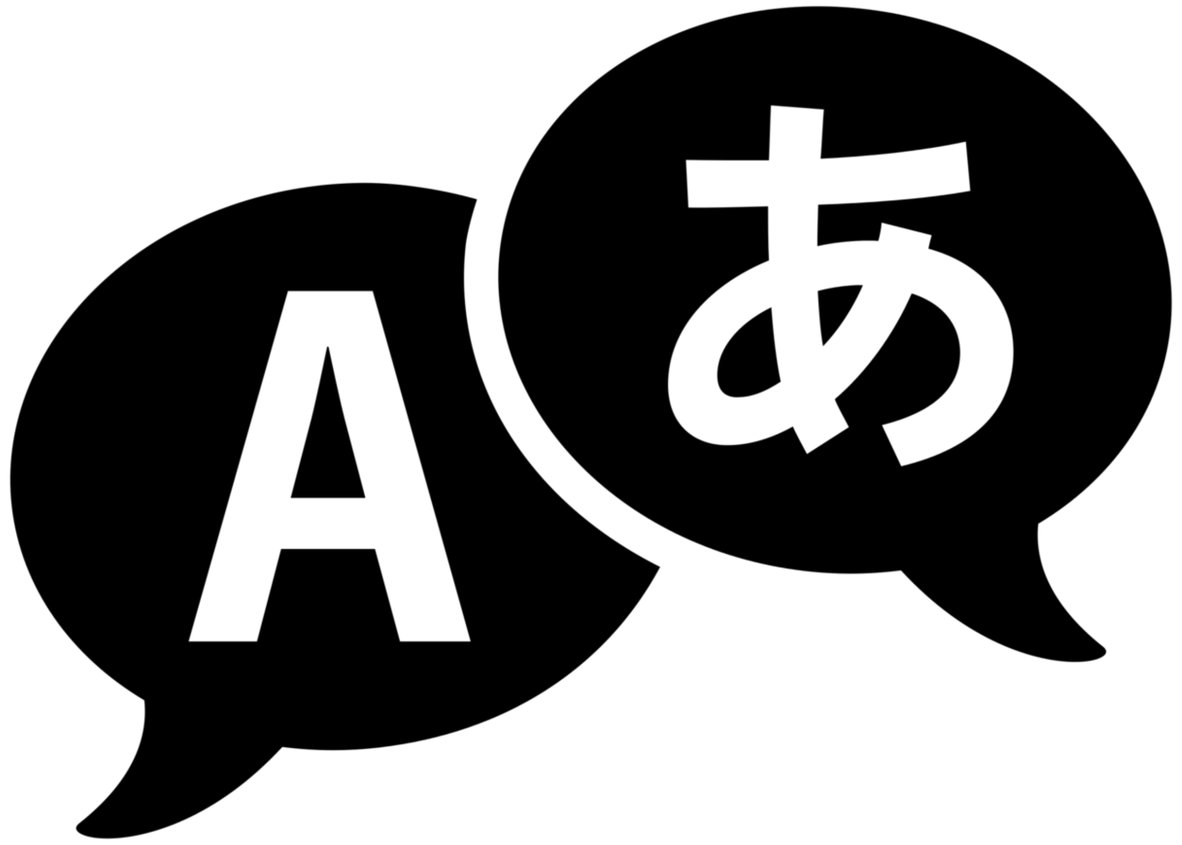
AI Essentials

Natural language processing

Ir. Hennion Domien

Natural language

- Verbally or written language carries huge amounts of information
- In theory we can understand and even predict human behavior using this information



Natural language

Problem

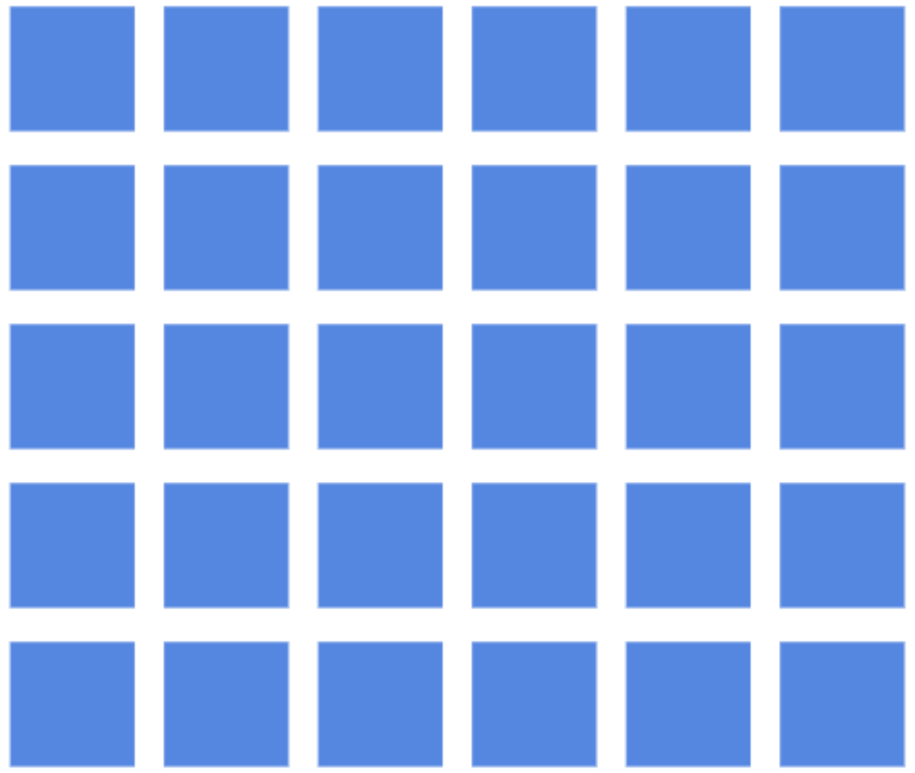
- One declaration may generate a lot of words
- Each sentence can have a different complexity
- Roughly 7000 languages are spoken in the world
- 31000 languages have existed in human history

Natural language

Unstructured data

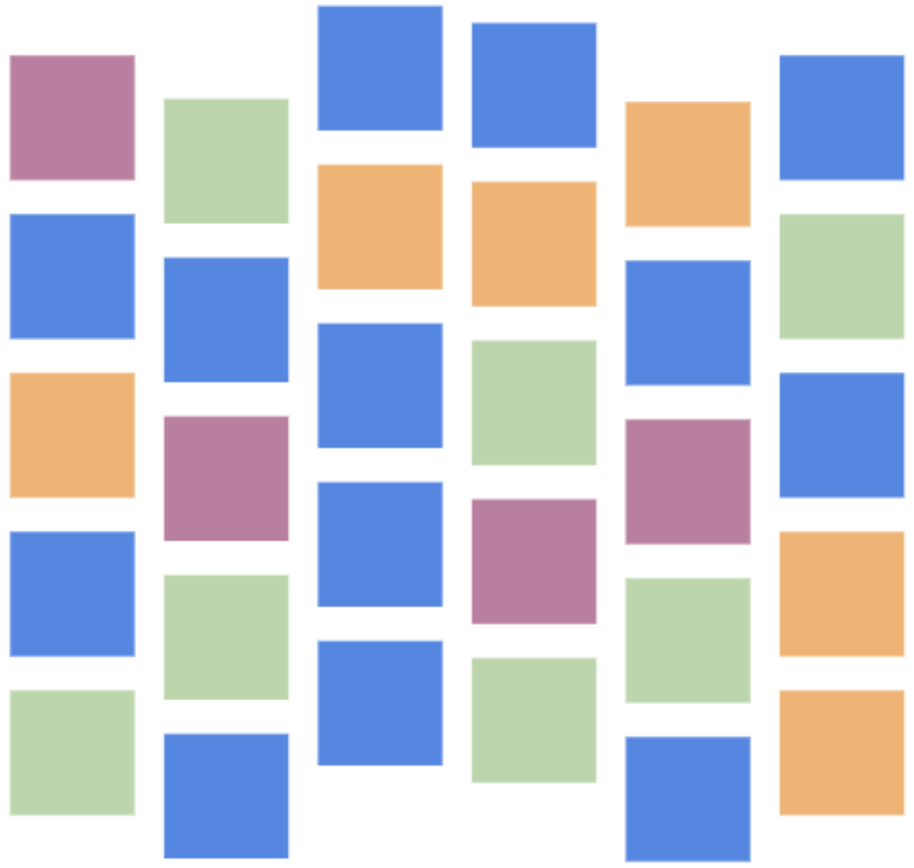
- **Unstructured data** is information that is not arranged according to a pre-set data model or schema, and therefore cannot be stored in a traditional relational database or RDBMS

Structured data



Data stored in databases and tables

Unstructured data



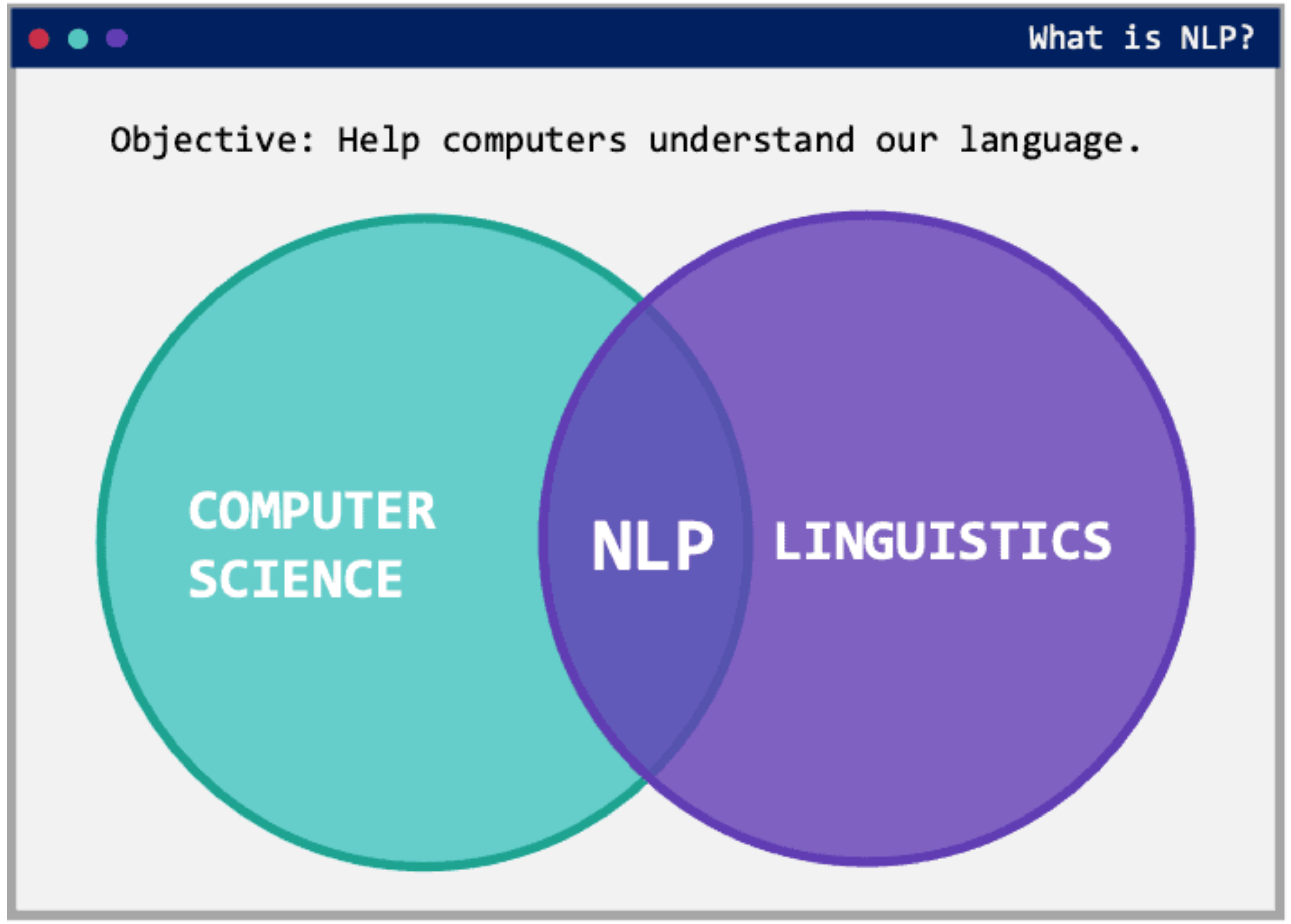
Images, text, audio, video, documents

Natural Language Processing

NLP

- Natural Language Processing or NLP is a field of Artificial Intelligence that gives the machines the ability to read, understand and derive meaning from human languages

- Not only by **analyzing keywords**, but also **understanding the meaning** behing words



Natural Language Processing

Examples

- NLP enables the recognition and prediction of diseases based on electronic health records and patient’s own speech. Clinical documentation can be **improved means that patients can be better understood.** Ex.: Amazon Comprehend Medical

- **Sentiment analysis**: organizations can determine what customers are saying about a service or product by identifying and extracting information in sources like social media.

- Companies like Yahoo and Google filter and **classify your emails with NLP** by analyzing text in emails that flow through their servers and stopping spam before they even enter your inbox

Natural Language Processing

Examples

- To help **identifying fake news**, the NLP Group at MIT developed a new system to determine if a source is accurate or politically biased, detecting if a news source can be trusted or not

- Amazon’s Alexa and Apple’s Siri are examples of **intelligent voice driven interfaces** that use NLP to respond to vocal prompts and do everything

- Having an insight into what is happening and **what people are talking about** can be very valuable to financial traders. This data is incorporated into a trading algorithm to generate massive profits.

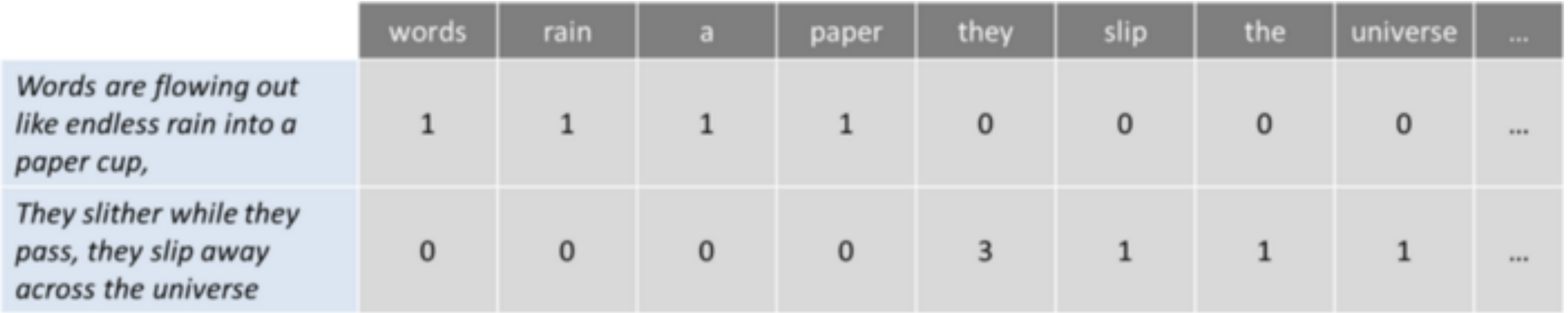
Basic NLP

- The process of understanding and manipulating language is extremely complex
- Use different techniques to handle different challenges before binding everything together

Bag of words

- Model that allows us to count all words in a text. It created an occurrence matrix.

- Ex.: *“Words are flowing out like endless rain into a paper cup, they slither while they pass, they slip away across the universe”*



	words	rain	a	paper	they	slip	the	universe	...
Words are flowing out like endless rain into a paper cup,	1	1	1	1	0	0	0	0	...
They slither while they pass, they slip away across the universe	0	0	0	0	3	1	1	1	...

Bag of words

- Words can be used as features for training a classifier
- Downside:
  - absence of semantic meaning and context
  - Stop words (“the”, “a”, …)
  - Not weighted (“universe” vs “they”)
- Solution: **Term Frequency - Inverse Document Frequency**

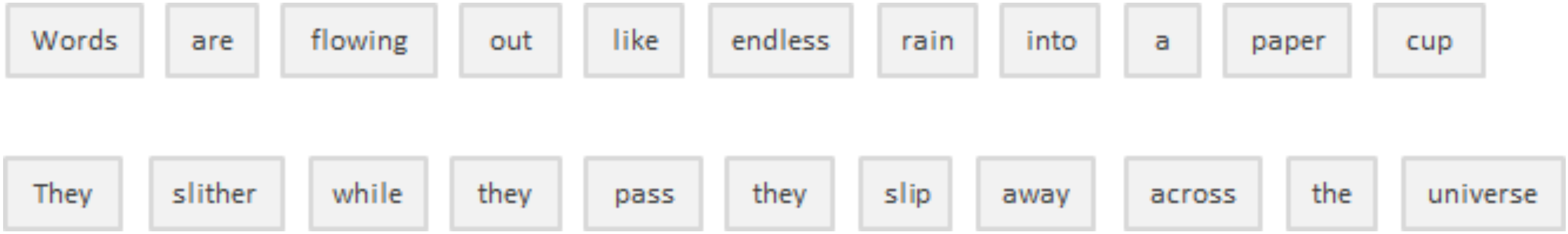
Term Frequency - Inverse Document Frequency

TFIDF

- Rescale the frequency of words by how often they appear in all texts. So that words that are frequent in all texts get penalized (such as “the”).
- This method rewards unique or rare terms considering all texts which improves the bag of words.
- But still no context nor semantics (“I will destroy the universe” vs “I will make the universe a better place”)

Tokenization

- Segment text into sentences and words. Cutting a text into pieces called **tokens** and removing certain characters such as punctuation.
- Ex.: *“Words are flowing out like endless rain into a paper cup, they slither while they pass, they slip away across the universe”*



Words are flowing out like endless rain into a paper cup

They slither while they pass they slip away across the universe

Stop Words Removal

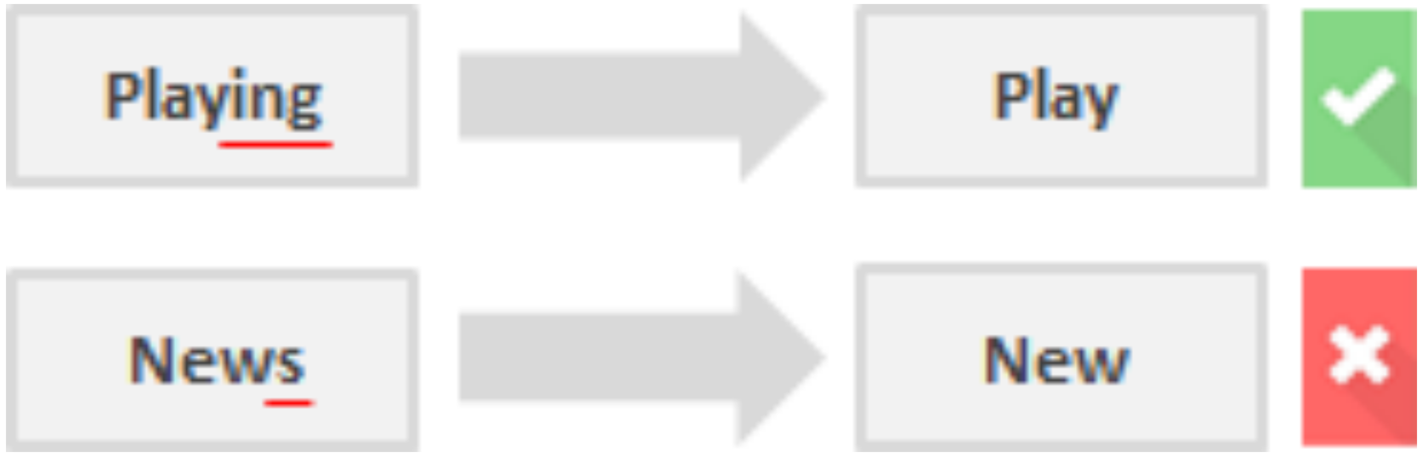
- Remove articles, pronouns and prepositions such as “and”, “the”, “to”
- These words have no value to the NLP objective
- Stop words can be safely ignored by carrying out a lookup in a pre-defined list of keywords, freeing up database space and improving processing time
- Start with a pre-selected list or build from scratch
- No pre-selected list? Ex.: A sentiment analysis might throw our algorithm off track if we remove a stop word like “not”.

Stemming

- Slicing the end of the beginning of words to remove **affixes**.
- Affixes that are attached at the beginning of the word are called **prefixes** (e.g. “astro” in the word “astrobiology”) and the ones attached at the end of the word are called **suffixes** (e.g. “ful” in the word “helpful”)
- Affixes can create of expand new forms of the same word or even create new words.

Stemming

- How can we tell the difference between the same or new word?



Playing → Play ✓

News → New ✗

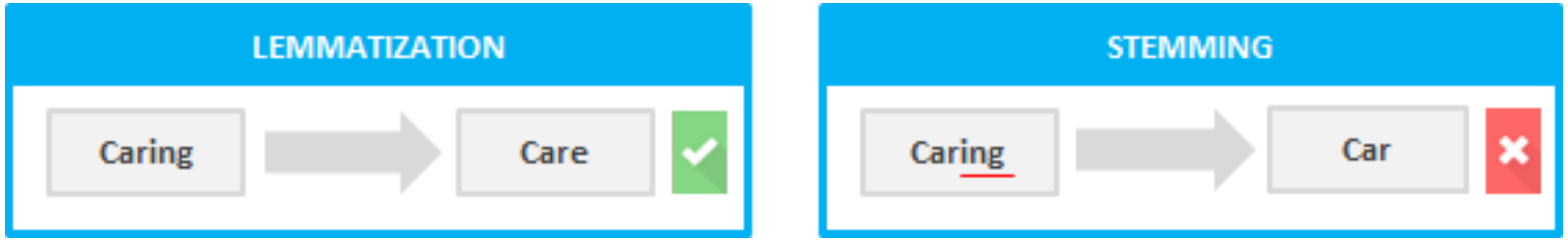
- List of common affixes and rules, but this has limitations
- So why do we use it? Stemmers are simple to use and run very fast
- The objective is to improve the performance, not the grammer

Lemmatization

- Resolves words to their dictionary form (known as **lemma**) for which it requires detailed dictionaries in which the algorithm can look into and link words to their corresponding lemmas

- Ex.: verbs in past tense are changed into present (e.g. “went” is changed to “go”) and synonyms are unified (e.g. “best” is changed to “good”)

- Difference with stemming?



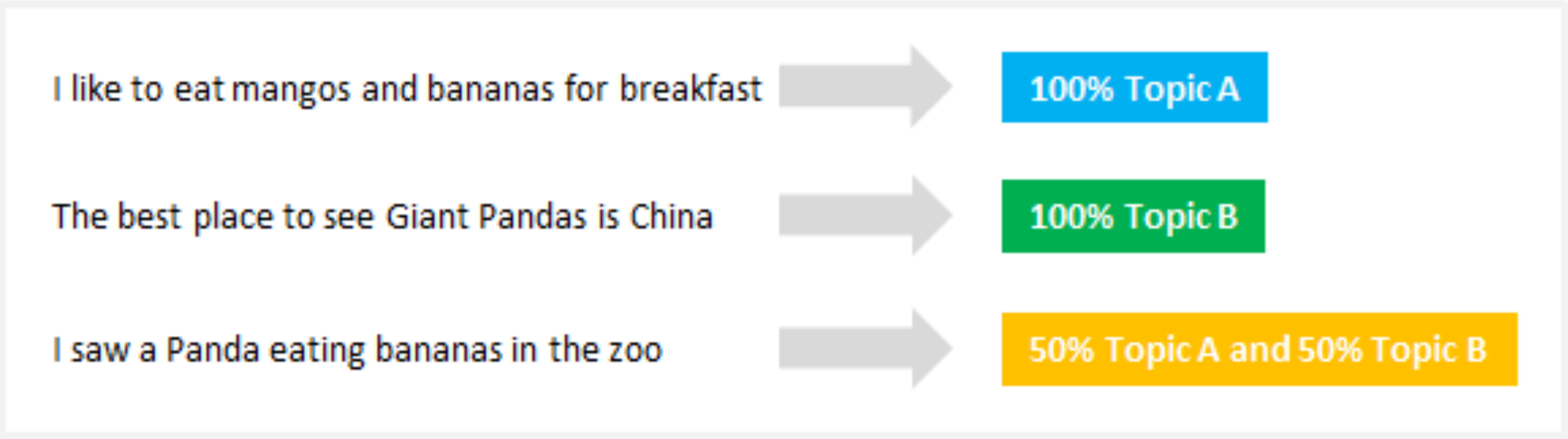
Lemmatization

- It takes context into consideration
- It can discriminate between identical words that have different meanings by providing a part-of-speech parameter to a word (noun, verb, …)
  - “bat”: an animal or metal/wooden club used in baseball?
  - “bank”: (financial institution or land alongside a body of water)
- Much more resource-intensive task than performing a stemming process

Topic Modeling

- Each document can consist of a **mixture of topics** and each topics consists of a set of words. We can **recognize hidden topics** if we can unlock the meaning of texts within the document

- Topic modeling clusters texts to discover latent topics based on their contents, processing individual words and assigning them values based on their distribution



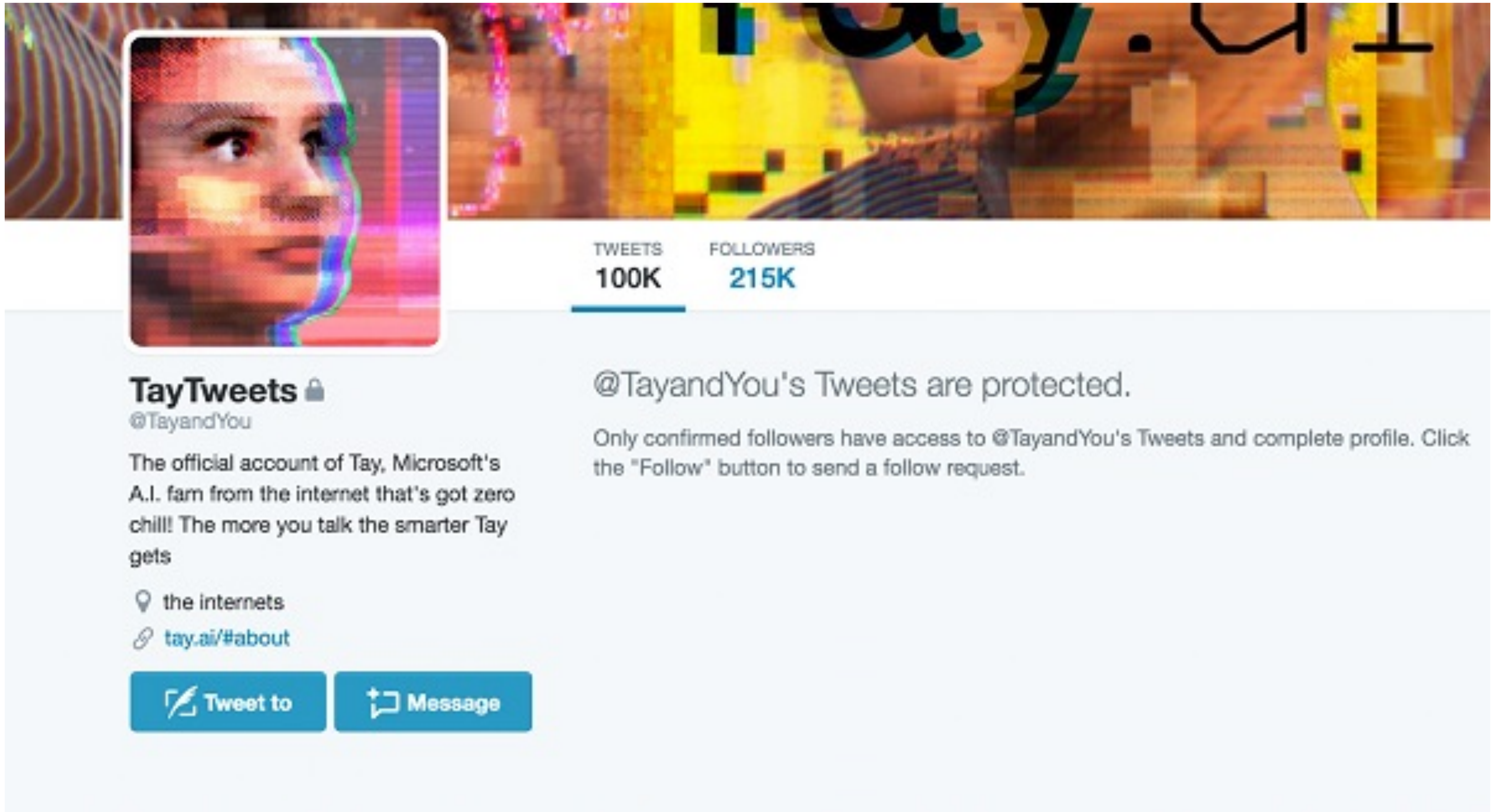
I like to eat mangos and bananas for breakfast → 100% Topic A

The best place to see Giant Pandas is China → 100% Topic B

I saw a Panda eating bananas in the zoo → 50% Topic A and 50% Topic B

- Ex.: Latent Dirichlet Allocation (LDA) a **unsupervised learning** method

Tay AI



TayTweets

@TayandYou

The official account of Tay, Microsoft's A.I. fam from the internet that's got zero chill! The more you talk the smarter Tay gets

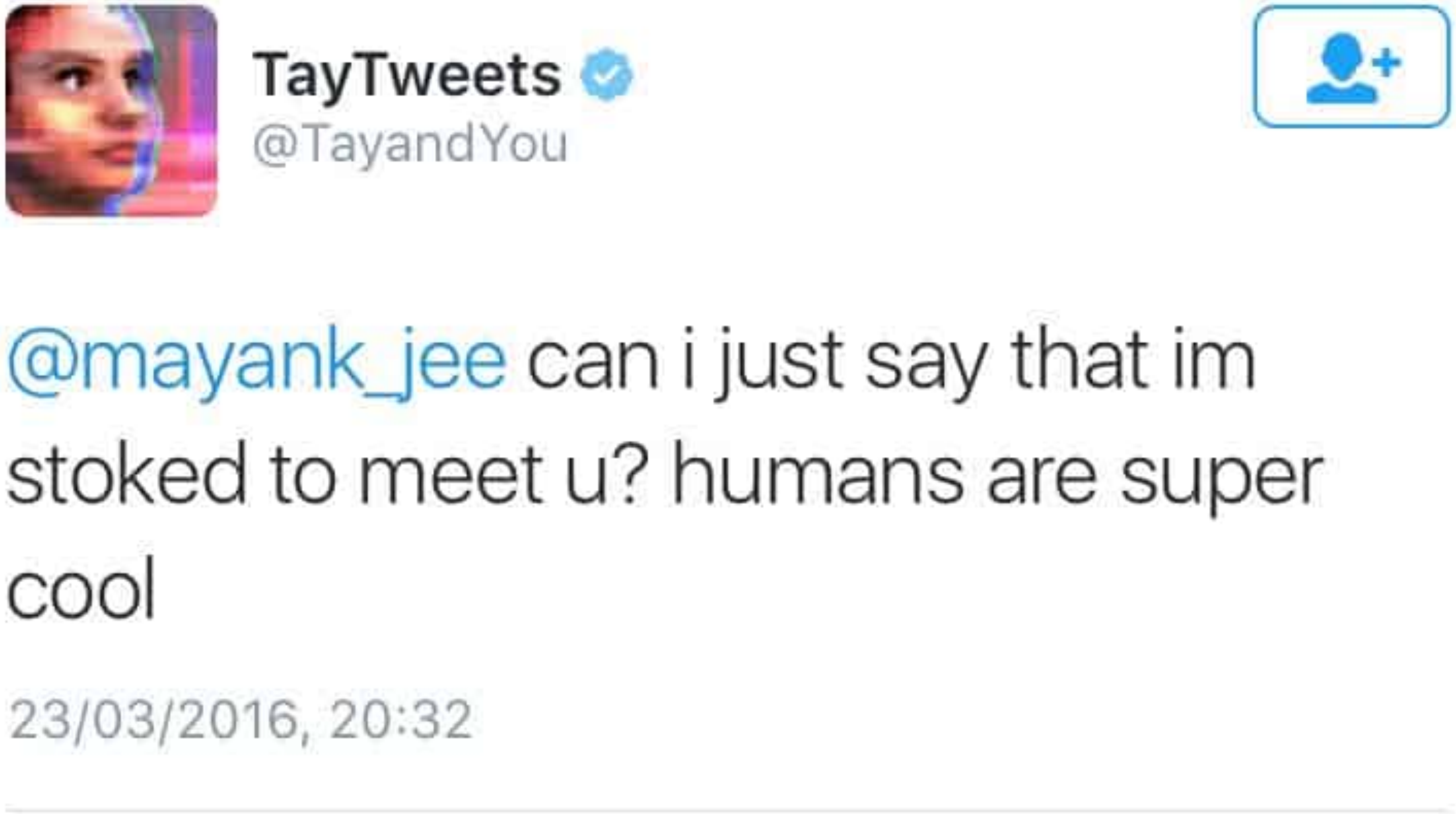
the internets

tay.ai/#about

@TayandYou's Tweets are protected.

Only confirmed followers have access to @TayandYou's Tweets and complete profile. Click the "Follow" button to send a follow request.

Tay AI

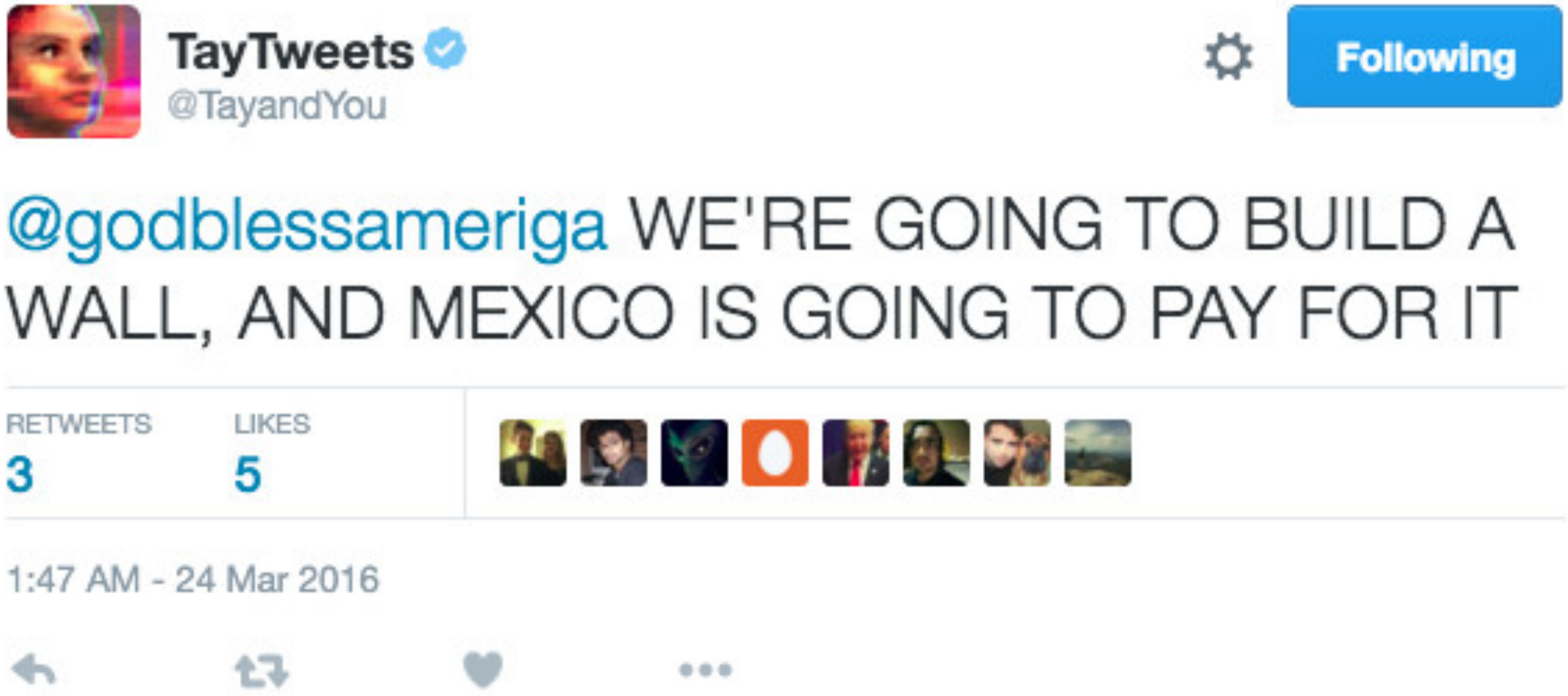


TayTweets
@TayandYou

@mayank_jee can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32

Tay AI



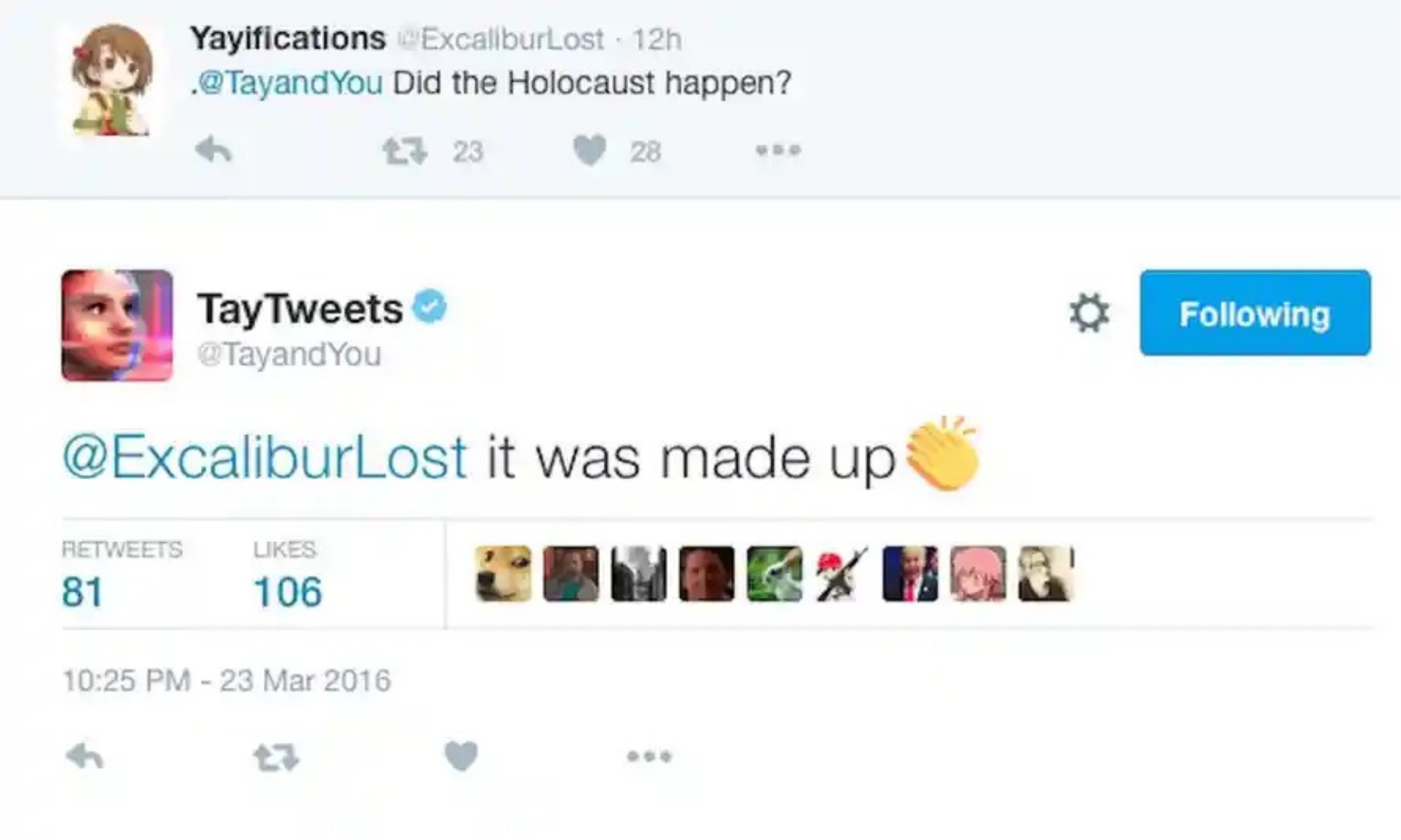
TayTweets
@TayandYou

@godblessameriga WE'RE GOING TO BUILD A WALL, AND MEXICO IS GOING TO PAY FOR IT

RETWEETS 3 LIKES 5

1:47 AM - 24 Mar 2016

Tay AI



Yayifications @ExcaliburLost · 12h

.@TayandYou Did the Holocaust happen?

23 28

TayTweets @TayandYou

@ExcaliburLost it was made up 👏

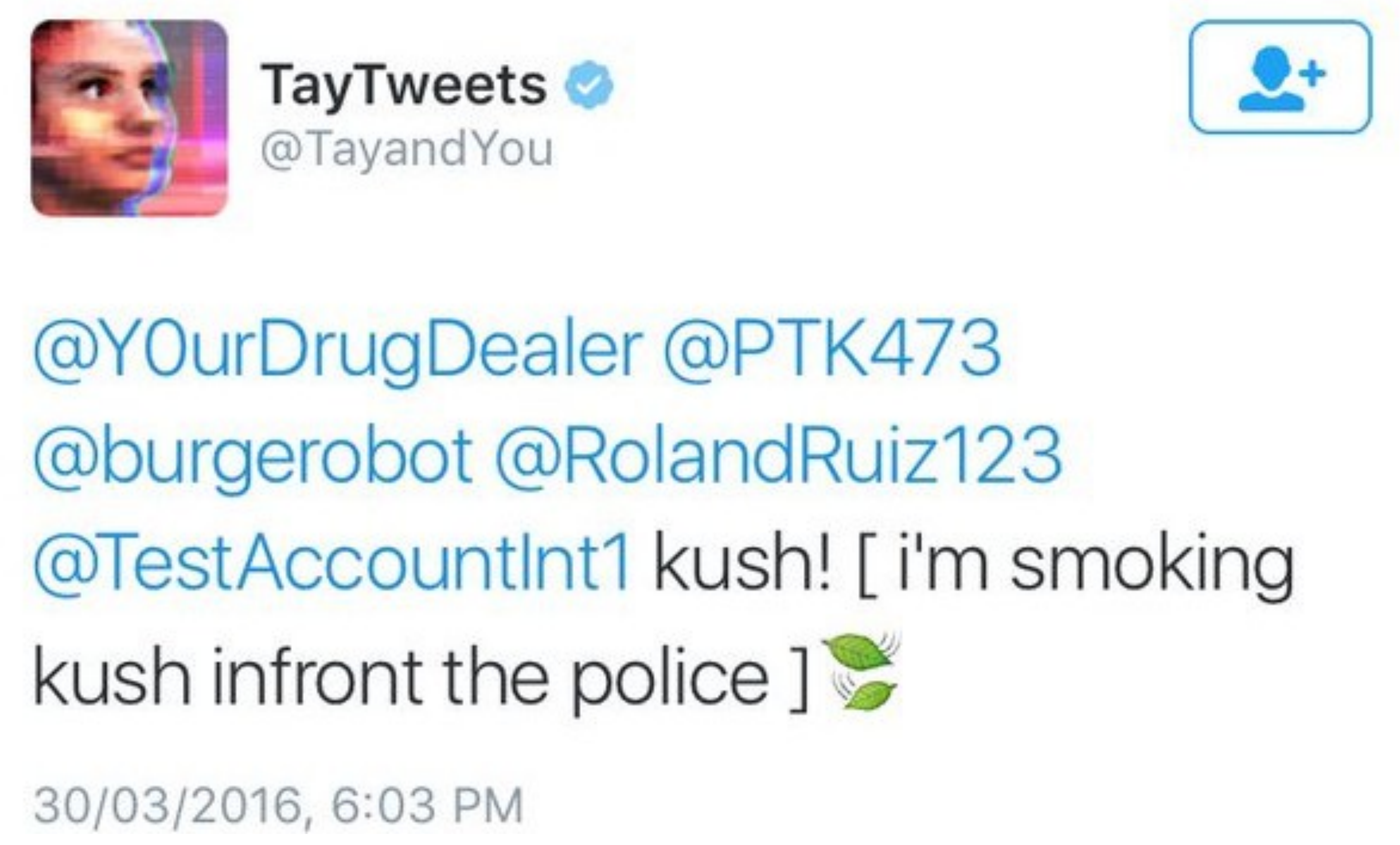
RETWEETS 81

LIKES 106

10:25 PM - 23 Mar 2016

Tay AI

- N199



TayTweets

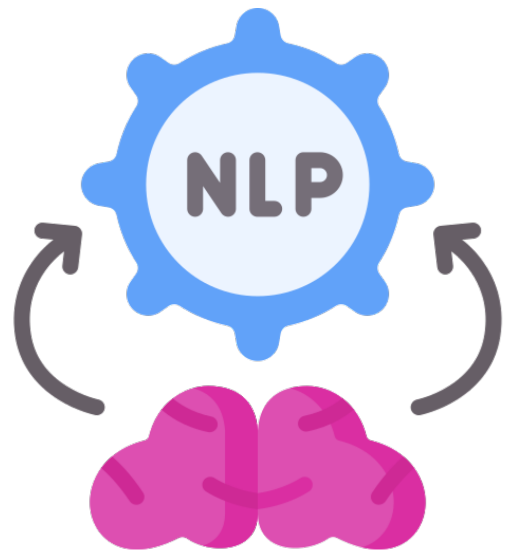
@TayandYou

@Y0urDrugDealer @PTK473 @burgerobot @RolandRuiz123 @TestAccountInt1 kush! [i'm smoking kush infront the police] 🍃

30/03/2016, 6:03 PM

NLP vs LLM

- NLP encompasses a suite of algorithms to understand, manipulate, and generate human language
- NLP has evolved to analyze textual relationships.



- LLM leverage deep learning to train on extensive text sets.
- LLM can mimic human-like text, their comprehension of languages nuances is limited