# Spooky Author Identification

By Cristina Iacob

# Goal

The aim of this project is to predict the author of some given paragraphs based on author's writing style. In particular, we will be analyzing texts from Edgar Allan Poe, Mary Shelley, and HP Lovecraft.

# Intended Audience

Anybody interested in an enhanced literature search tool to find the (most likely) author of a preferred piece, as well as other writers with similar literary style

Ex: an online bookstore which want to improve user's experience, also resulting in increased sales

# Data Description

The dataset contains text from works of fiction written by Edgar Allan Poe, HP Lovecraft and Mary Shelley.

***File descriptions***

- train.csv - the training set - 19,579 samples
- test.csv - the test set - 8,392 samples (no labels)

***Data fields***

- id - a unique identifier for each sentence
- text - some text written by one of the authors
- author - the author of the sentence (EAP: Edgar Allan Poe, HPL: H.P. Lovecraft; MWS: Mary Wollstonecraft Shelley)

Data source:

https://www.kaggle.com/c/spooky-author-identification

# Our Authors

**Edgar Allen Poe : American writer who wrote poetry and short stories that revolved around tales of mystery and the grisly and the grim. Arguably his most famous work is the poem - "The Raven" and he is also widely considered the pioneer of the genre of the detective fiction.**

**H.P. Lovecraft : Best known for authoring works of horror fiction, the stories that he is most celebrated for revolve around the fictional mythology of the infamous creature "Cthulhu" - a hybrid chimera mix of Octopus head and humanoid body with wings on the back.**

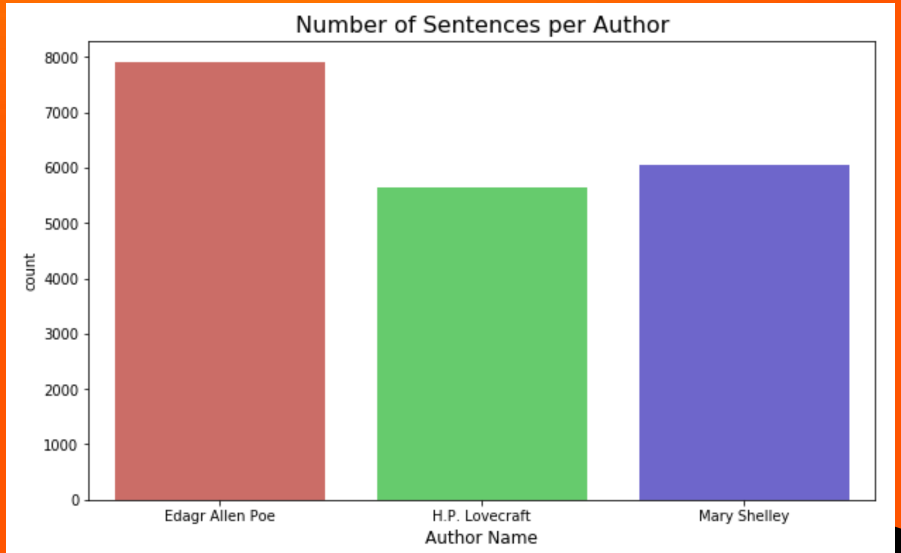

**Mary Wollstonecraft Shelley : Novelist, dramatist, travel-writer, biographer. She is most celebrated for the classic tale of Frankenstein where the scientist Frankenstein a.k.a "The Modern Prometheus" creates the Monster that comes to be associated with his name.**

# Visualizing data

| | id | text | author |
|---|---|---|---|
| 0 | id26305 | This process, however, afforded me no means of... | EAP |
| 1 | id17569 | It never once occurred to me that the fumbling... | HPL |
| 2 | id11008 | In his left hand was a gold snuff box, from wh... | EAP |
| 3 | id27763 | How lovely is spring As we looked from Windsor... | MWS |
| 4 | id12958 | Finding nothing else, not even gold, the Super... | HPL |

**Author statistic :**

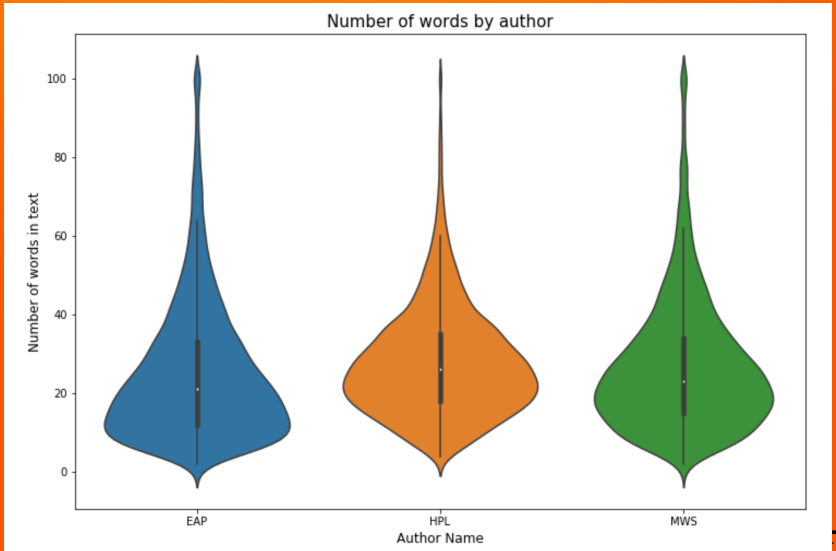
- EAP – 7900 (40%)
- MWS – 6044 (31%)
- HPL – 5635 (29%)



# Visualizing data (cont'd)

| | author | num_words | num_unique_words | num_chars | num_stopwords | num_punctuations | num_words_upper | mean_word_len |
|---|---|---|---|---|---|---|---|---|
| **0** | EAP | 41 | 35 | 231 | 19 | 7 | 2 | 4.658537 |
| **1** | HPL | 14 | 14 | 71 | 8 | 1 | 0 | 4.142857 |
| **2** | EAP | 36 | 32 | 200 | 16 | 5 | 0 | 4.583333 |
| **3** | MWS | 34 | 32 | 206 | 13 | 4 | 0 | 5.088235 |
| **4** | HPL | 27 | 25 | 174 | 11 | 4 | 0 | 5.481481 |

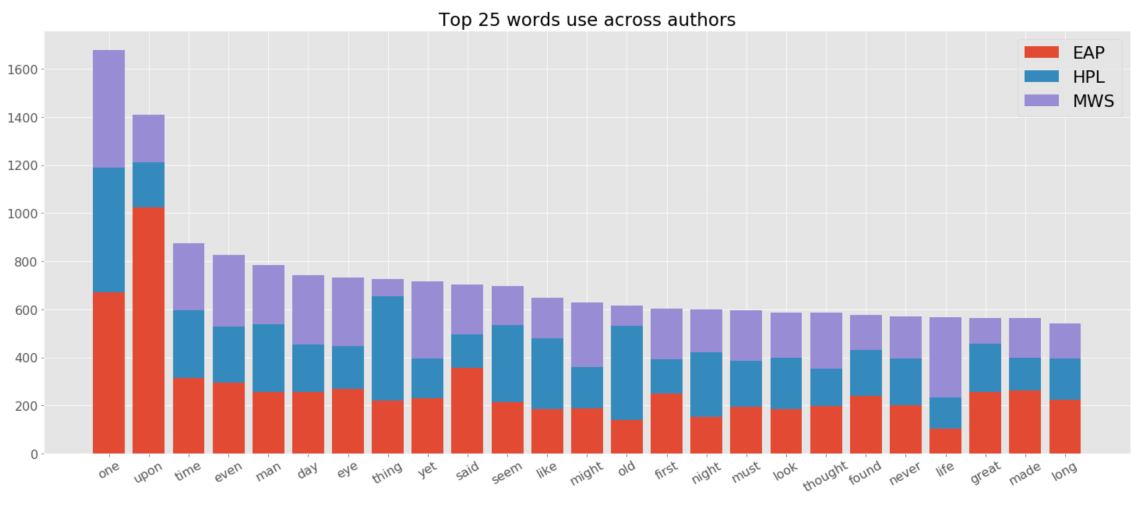
## Numerical stats



# Visualizing data (cont'd)

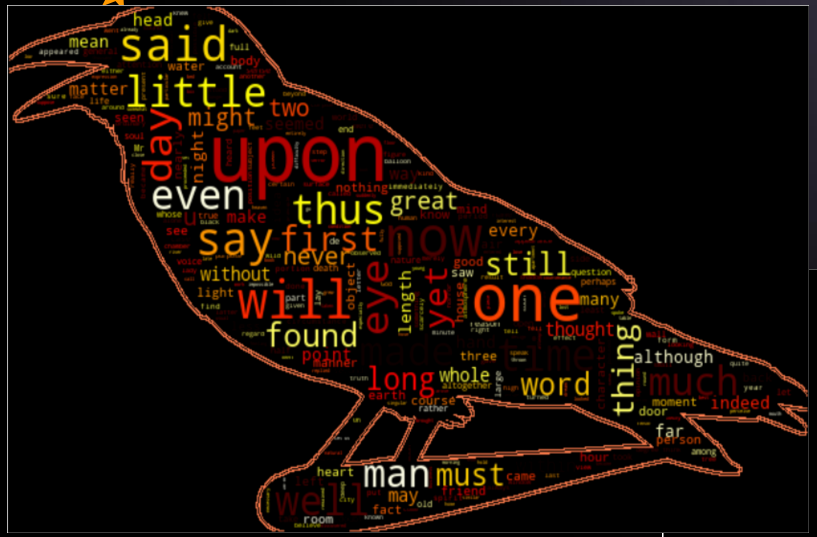
## Common Words

| | id | text | author | tokenized | stemmed |
|---|---|---|---|---|---|
| 0 | id26305 | This process, however, afforded me no means of... | EAP | [This, process, ,, however, ,, afforded, me, n... | [process, howev, afford, mean, ascertain, dime... |
| 1 | id17569 | It never once occurred to me that the fumbling... | HPL | [It, never, once, occurred, to, me, that, the,... | [never, occur, fumbl, might, mere, mistak] |
| 2 | id11008 | In his left hand was a gold snuff box, from wh... | EAP | [In, his, left, hand, was, a, gold, snuff, box... | [left, hand, gold, snuff, box, caper, hill, cu... |
| 3 | id27763 | How lovely is spring As we looked from Windsor... | MWS | [How, lovely, is, spring, As, we, looked, from... | [love, spring, look, windsor, terrac, sixteen,... |
| 4 | id12958 | Finding nothing else, not even gold, the Super... | HPL | [Finding, nothing, else, ,, not, even, gold, ,... | [find, noth, els, even, gold, superintend, aba... |



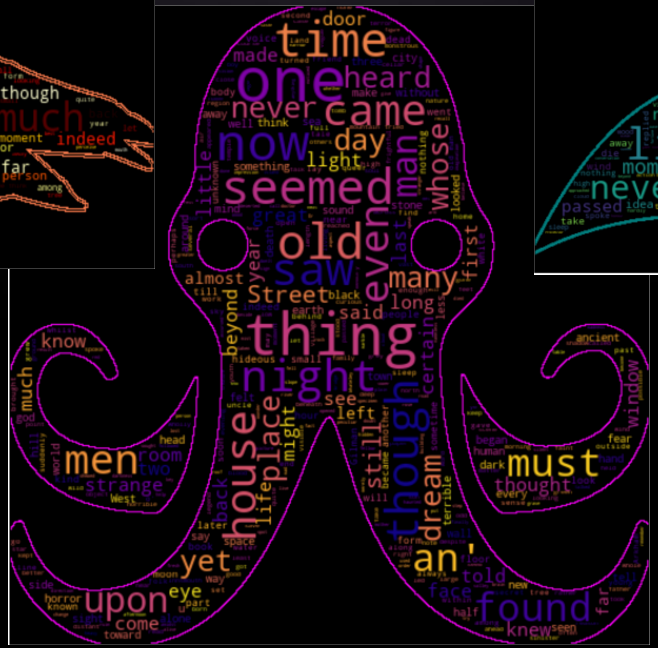

| | ALL | EAP | HPL | MWS |
|---|---|---|---|---|
| one | 1677 | 672 | 516 | 489 |
| upon | 1411 | 1025 | 186 | 200 |
| time | 874 | 315 | 281 | 278 |
| even | 828 | 296 | 234 | 298 |
| man | 786 | 258 | 279 | 249 |
| day | 743 | 258 | 197 | 288 |
| eye | 732 | 270 | 176 | 286 |
| thing | 725 | 221 | 433 | 71 |
| yet | 715 | 232 | 165 | 318 |
| said | 704 | 356 | 140 | 208 |
| seem | 696 | 214 | 322 | 160 |
| like | 649 | 185 | 296 | 168 |
| might | 629 | 188 | 172 | 269 |
| old | 616 | 139 | 392 | 85 |
| first | 603 | 250 | 142 | 211 |
| night | 600 | 154 | 268 | 178 |
| must | 597 | 196 | 189 | 212 |
| look | 588 | 186 | 213 | 189 |
| thought | 588 | 198 | 157 | 233 |
| found | 576 | 239 | 191 | 146 |
| never | 570 | 202 | 193 | 175 |
| life | 569 | 105 | 130 | 334 |
| great | 565 | 255 | 203 | 107 |
| made | 565 | 263 | 136 | 166 |
| long | 541 | 224 | 171 | 146 |

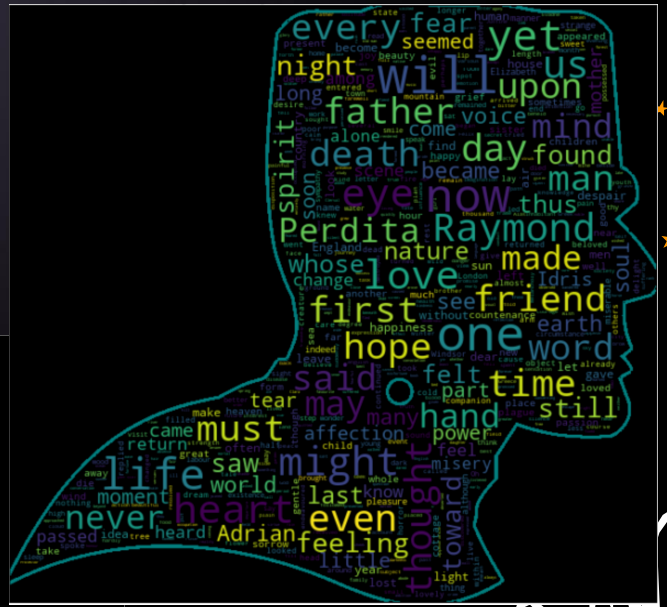
# Wordcloud




EAP

HPL




MWS

# Processed text

| | id | text | author | tokenized | tagged | stemmed |
|---|---|---|---|---|---|---|
| 0 | id26305 | This process, however, afforded me no means of... | EAP | [This, process, ,, however, ,, afforded, me, n... | [DT, NN, ,, RB, ,, VBD, PRP, DT, NNS, IN, VBG,... | [process, howev, afford, mean, ascertain, dime... |
| 1 | id17569 | It never once occurred to me that the fumbling... | HPL | [It, never, once, occurred, to, me, that, the,... | [PRP, RB, RB, VBD, TO, PRP, IN, DT, NN, MD, VB... | [never, occur, fumbl, might, mere, mistak] |
| 2 | id11008 | In his left hand was a gold snuff box, from wh... | EAP | [In, his, left, hand, was, a, gold, snuff, box... | [IN, PRP$, JJ, NN, VBD, DT, JJ, NN, NN, ,, IN,... | [left, hand, gold, snuff, box, caper, hill, cu... |
| 3 | id27763 | How lovely is spring As we looked from Windsor... | MWS | [How, lovely, is, spring, As, we, looked, from... | [WRB, RB, VBZ, JJ, IN, PRP, VBD, IN, NNP, NNP,... | [love, spring, look, windsor, terrac, sixteen,... |
| 4 | id12958 | Finding nothing else, not even gold, the Super... | HPL | [Finding, nothing, else, ,, not, even, gold, ,... | [VBG, NN, RB, ,, RB, RB, NN, ,, DT, NNP, VBD, ... | [find, noth, els, even, gold, superintend, aba... |

Tokenizing, lemmatizing, stemming, removing stop words and punctuation, tagging parts of speech

| | count | explanation | percentage |
|---|---|---|---|
| NN | 84783 | noun, common, singular or mass | 14% |
| IN | 71911 | preposition or conjunction, subordinating | 12% |
| DT | 59511 | determiner | 10% |
| JJ | 38616 | adjective or numeral, ordinal | 6% |
| , | 38220 | comma | 6% |
| PRP | 34441 | pronoun, personal | 6% |
| VBD | 34140 | verb, past tense | 6% |
| RB | 29458 | adverb | 5% |
| CC | 23640 | conjunction, coordinating | 4% |
| NNS | 23429 | noun, common, plural | 4% |
| . | 20217 | sentence terminator | 3% |
| VB | 17049 | verb, base form | 3% |
| VBN | 15715 | verb, past participle | 3% |
| PRP$ | 15446 | pronoun, possessive | 3% |
| NNP | 14952 | noun, proper, singular | 3% |
| TO | 12838 | "to" as preposition or infinitive marker | 2% |
| VBG | 7927 | verb, present participle or gerund | 1% |
| MD | 6835 | modal auxiliary | 1% |
| VBP | 6769 | verb, present tense, not 3rd person singular | 1% |
| : | 5790 | colon or ellipsis | 1% |

# Preprocessing & Modeling

**Data**: randomly split : 80% for Training and 20% for Testing

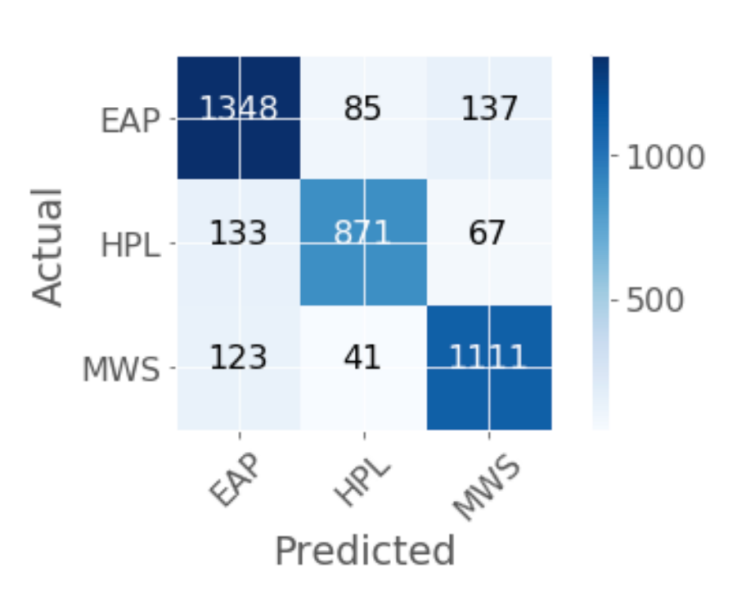
1. Text was processed and encoded as BoW (output features containing unique words and a count of occurrences)
2. The text was “normalized” (using tf-idf transformer)
3. Applied ML Classification algorithms

## **Selected Models**:

- Logistic Regression
- K Nearest Neighbors Classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- Multinomial NB

# Results

The winner is Multinomial NB with an 85% accuracy




3916 total excerpt:

- EAP: 1604 (1348) (256)
- HPL: 997 (871) (126)
- MS: 1317 (1111) (204)

# Conclusions and Next Steps

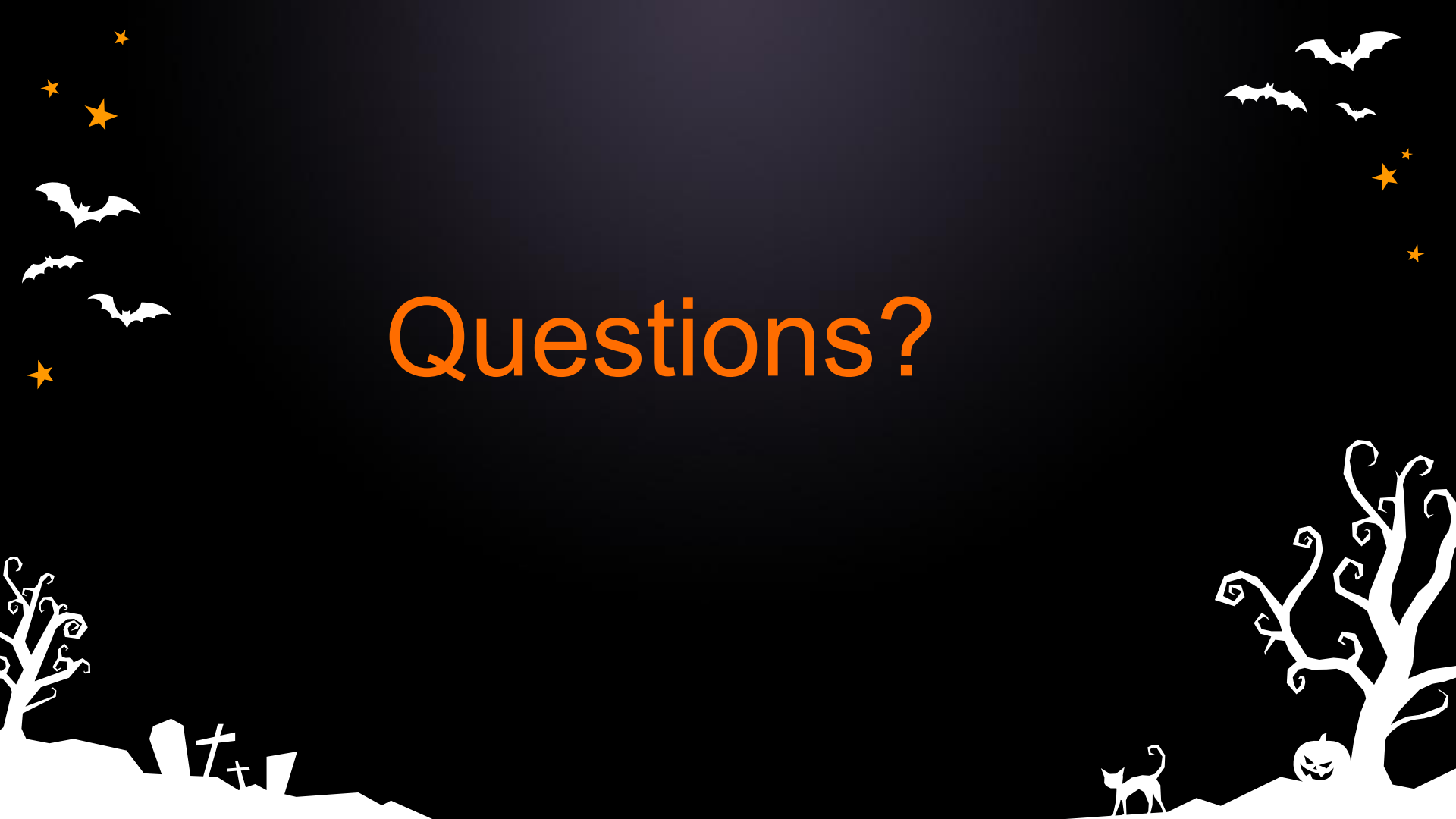
We can improve the results by:

- Using other ML classification algorithms
- Using neural networks models (such CNN)
- Use a vocabulary of 3-grams containing different part of speech
- Use a different method for text encoding (such as word2vec)
- We can find more insights from the analysis and visualization of data

Different analysis :

Can do a sentiment analysis; gender analysis; usage of stopwords, unique words, punctuations etc.

# Questions?



Thank you!