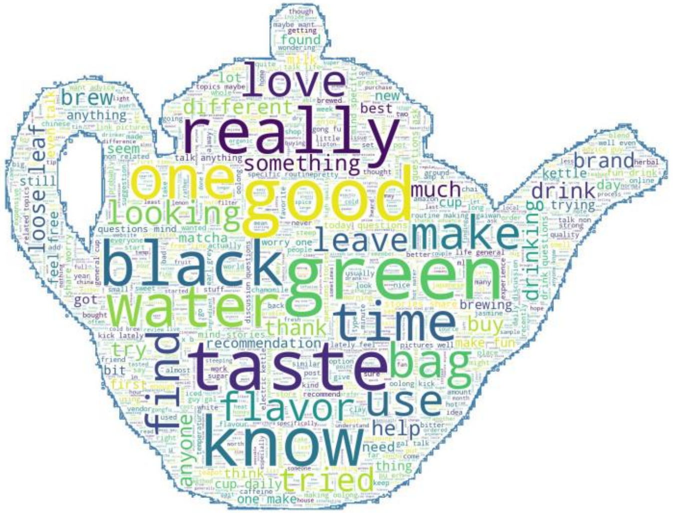
Web APIs and NLP Classification



Objective & Aim

- Objective
- A beverage company wants to know if people in an area posts more on social media about coffee or tea.

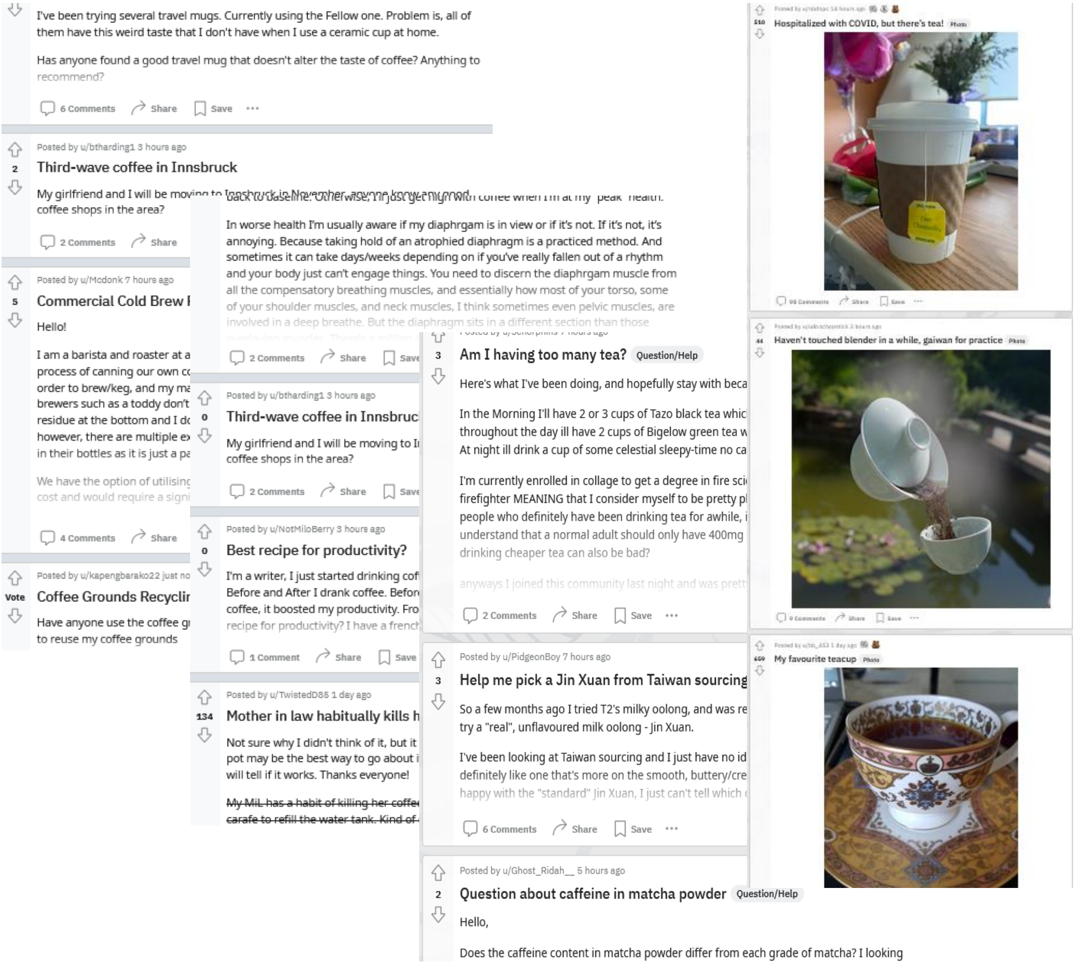
- Aim
- A classification model is needed to classify posts into either coffee or tea.

Content

- Data Gathering
- Data Cleaning
- Naïve Bayes Model
- Alternative Models
- Conclusions and Recommendations

Data Collection

- Using Pushshift's API, posts about coffee and tea were collected.
- Empty, removed or deleted posts were removed.
- Duplicates were removed.



I've been trying several travel mugs. Currently using the Fellow one. Problem is, all of them have this weird taste that I don't have when I use a ceramic cup at home.

Has anyone found a good travel mug that doesn't alter the taste of coffee? Anything to recommend?

6 Comments Share Save ···

Posted by u/btharding1 3 hours ago
2
Third-wave coffee in Innsbruck

My girlfriend and I will be moving to Innsbruck in November, anyone know any good coffee shops in the area?

2 Comments Share

Posted by u/Mcdonk 7 hours ago
5
Commercial Cold Brew

Hello!

I am a barista and roaster at a
process of canning our own co
order to brew/keg, and my ma
brewers such as a toddy don't
residue at the bottom and I do
however, there are multiple ex
in their bottles as it is just a pa

We have the option of utilising
cost and would require a signi

4 Comments Share

Posted by u/kapengbarako22 just no
Vote
Coffee Grounds Recyclin

Have anyone use the coffee gr
to reuse my coffee grounds

back to baseline. Otherwise, I'll just get high with coffee when I'm at my "peak" health.

In worse health I'm usually aware if my diaphrgam is in view or if it's not. If it's not, it's annoying. Because taking hold of an atrophied diaphragm is a practiced method. And sometimes it can take days/weeks depending on if you've really fallen out of a rhythm and your body just can't engage things. You need to discern the diaphrgam muscle from all the compensatory breathing muscles, and essentially how most of your torso, some of your shoulder muscles, and neck muscles, I think sometimes even pelvic muscles, are involved in a deep breathe. But the diaphragm sits in a different section than those

2 Comments Share Save

Posted by u/btharding1 3 hours ago
0
Third-wave coffee in Innsbruc

My girlfriend and I will be moving to In
coffee shops in the area?

2 Comments Share Save

Posted by u/NotMiloBerry 3 hours ago
0
Best recipe for productivity?

I'm a writer, I just started drinking cof
Before and After I drank coffee. Befor
coffee, it boosted my productivity. Fro
recipe for productivity? I have a french

1 Comment Share Save

Posted by u/TwistedD85 1 day ago
134
Mother in law habitually kills h

Not sure why I didn't think of it, but it
pot may be the best way to go about i
will tell if it works. Thanks everyone!

~~My MiL has a habit of killing her coffee~~
~~carafe to refill the water tank. Kind of~~

3
Am I having too many tea? Question/Help

Here's what I've been doing, and hopefully stay with beca

In the Morning I'll have 2 or 3 cups of Tazo black tea whic
throughout the day ill have 2 cups of Bigelow green tea w
At night ill drink a cup of some celestial sleepy-time no ca

I'm currently enrolled in collage to get a degree in fire sci
firefighter MEANING that I consider myself to be pretty pl
people who definitely have been drinking tea for awhile, i
understand that a normal adult should only have 400mg
drinking cheaper tea can also be bad?

anyways I joined this community last night and was prett

2 Comments Share Save ···

Posted by u/PidgeonBoy 7 hours ago
3
Help me pick a Jin Xuan from Taiwan sourcing

So a few months ago I tried T2's milky oolong, and was re
try a "real", unflavoured milk oolong - Jin Xuan.

I've been looking at Taiwan sourcing and I just have no id
definitely like one that's more on the smooth, buttery/cre
happy with the "standard" Jin Xuan, I just can't tell which

6 Comments Share Save ···

Posted by u/Ghost_Ridah__ 5 hours ago
2
Question about caffeine in matcha powder Question/Help

Hello,

Does the caffeine content in matcha powder differ from each grade of matcha? I looking

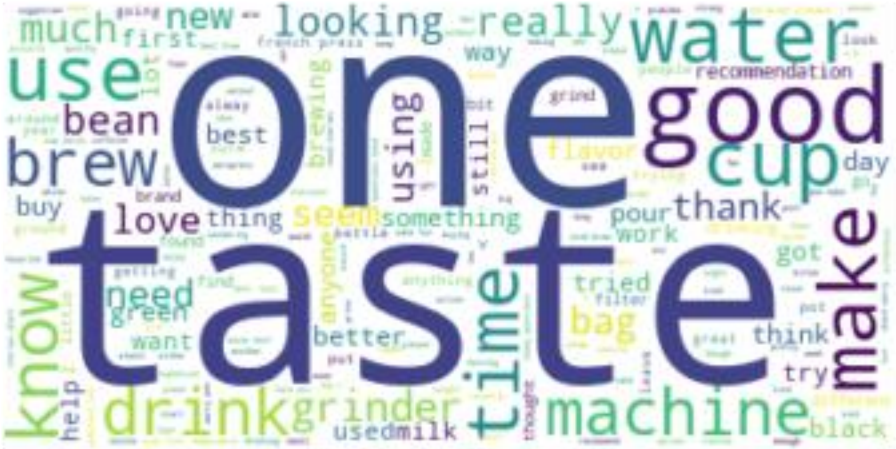
510
Hospitalized with COVID, but there's tea! Photo

44
Haven't touched blender in a while, gaiwan for practice Photo

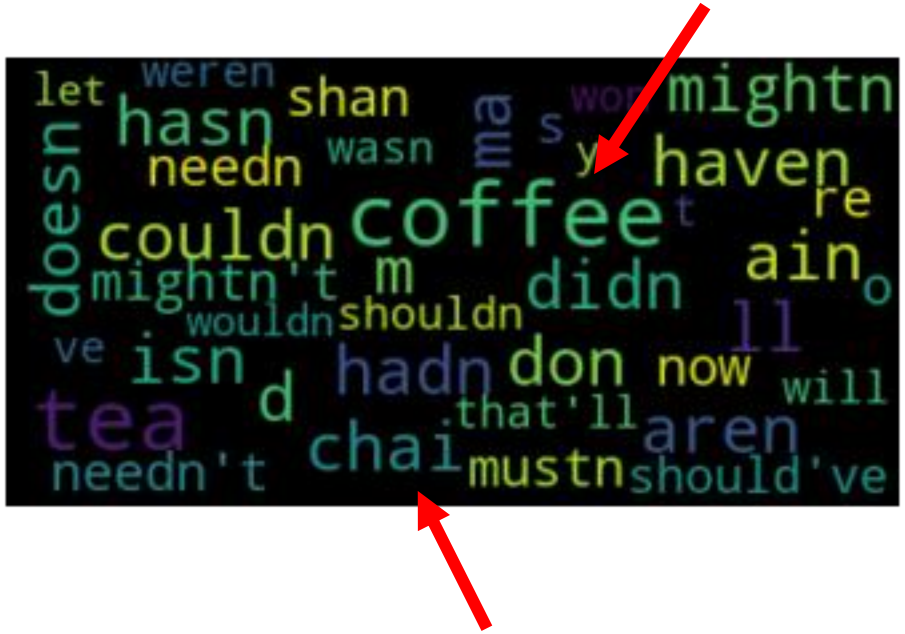
659
My favourite teacup Photo

Data Cleaning

Original text



Removal of stop words



Data Cleaning

Original text

Get a fine italian expresso, melt a little square of dark chocolate in it (at least 70% cocoa) and a dash of cinemon. Nothing more!\n\nAnd then, Like Homer said, "It\'s like a party in my mouth and everyone\'s invited!"Just for real cofee lovers!

After data cleaning

get fine italian melt little square dark chocolate least cocoa dash cinemon nothing like homer said like party mouth everyone invited real cofee lovers

Lemmatization and Stemming

- **Cleaned sample**
- using starbucks beans grinding size tamping twice get oz secondshelp brevle dual temp machine

- **Lemmatization**
- using starbucks bean grinding size tamping twice get oz secondshelp brevle dual temp machine

- **Stemming**
- use starbuck bean grind size tamp twice get oz secondshelp brevl dual temp machin

Naïve Bayes Model

- MultinomialNB
- It is easy to implement.
- One of the most popular supervised learning classification.

- Metrics:
- Accuracy score – emphasis on correct positive predictions.
- f1-score – emphasis on Type 1 and Type 2 errors.

Naïve Bayes Model

- No differences in the scores when using cleaned, lemmatized and stemmed datasets.

Dataset	Train score	Accuracy	f1-score
Cleaned	0.893	0.871	0.879
Lemmatized	0.895	0.872	0.879
Stemmed	0.897	0.865	0.870

Naïve Bayes Model – Feature Importance

- ‘bean’
- ‘grind’, ‘burr', ‘roaster’,

- ‘machines’
- ‘breville', ‘baratza', ‘jx’, ‘zpresso’

- Method
- ‘moka’, ’drip’, ‘barista’

Naïve Bayes Model – Feature Importance

- Types
- 'chinese', 'sencha', 'yunnan', 'oolong', 'chamomile', ‘pu’, ‘erh’, ‘gong’, ‘fu’

- Device
- 'gaiwan’

- Other
- ‘leaf’, 'sourcing’,
- ‘worry’, ‘stories’, ‘topics’, ‘august’, ‘gal’

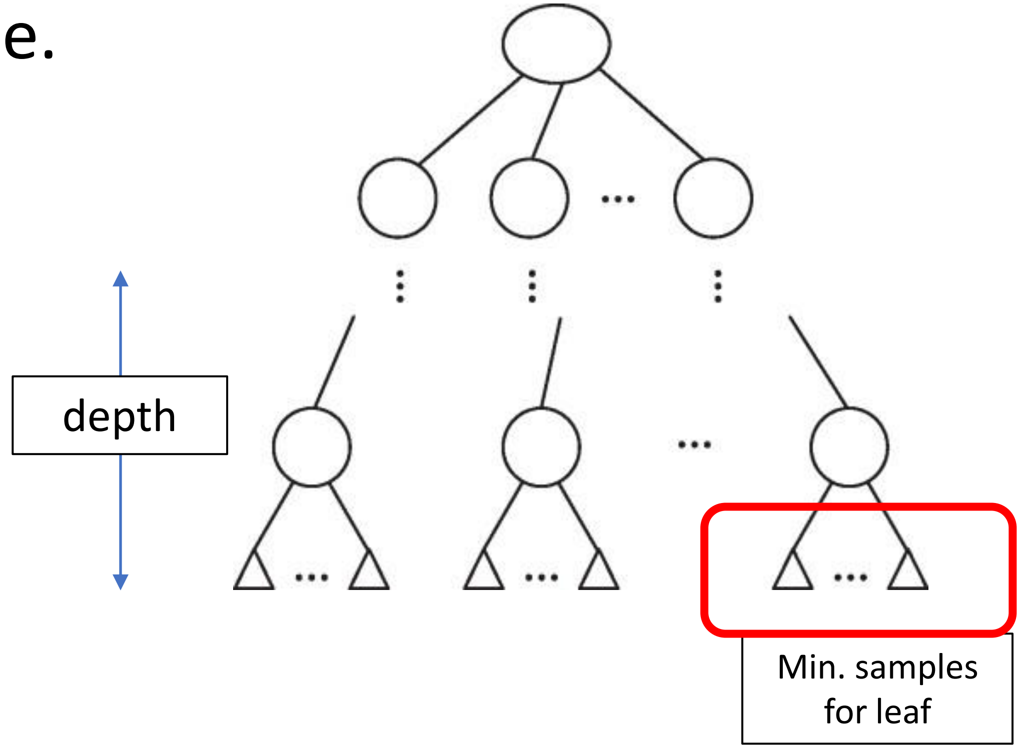
Spam !!

- Not dropped during cleaning.
- They were properly classified as ‘tea’.

drinking today questions mind stories share worry one make fun drink questions ask also talk anything else mind specific routine making oolong kick lately feel free link pictures well even talk non related topics maybe want advice guy gal talk life general cup daily discussion questions stories September
drinking today questions mind stories share worry one make fun drink questions ask also talk anything else mind specific routine making oolong kick lately feel free link pictures well even talk non related topics maybe want advice guy gal talk life general cup daily discussion questions stories September
drinking today questions mind stories share worry one make fun drink questions ask also talk anything else mind specific routine making oolong kick lately feel free link pictures well even talk non related topics maybe want advice guy gal talk life general cup daily discussion questions stories September
drinking today questions mind stories share worry one make fun drink questions ask also talk anything else mind specific routine making oolong kick lately feel free link pictures well even talk non related topics maybe want advice guy gal talk life general cup daily discussion questions stories September
drinking today questions mind stories share worry one make fun drink questions ask also talk anything else mind specific routine making oolong kick lately feel free link pictures well even talk non related topics maybe want advice guy gal talk life general cup daily discussion questions stories September
drinking today questions mind stories share worry one make fun drink questions ask also talk anything else mind specific routine making oolong kick lately feel free link pictures well even talk non related topics maybe want advice guy gal talk life general cup daily discussion questions stories September
drinking today questions mind stories share worry one make fun drink questions ask also talk anything else mind specific routine making oolong kick lately feel free link pictures well even talk non related topics maybe want advice guy gal talk life general cup daily discussion questions stories September
drinking today questions mind stories share worry one make fun drink questions ask also talk anything else mind specific routine making oolong kick lately feel free link pictures well even talk

Alternative Models – Decision Tree Classifier

- One decision tree, one outcome.
- Not recommended.
- Tendency to overfit.
- Hyperparameter tuning:
  - Tree depth
  - Min sample for split
  - Min samples for leaf



Random Forest vs Extra Trees

- Hyperparameter tuning:
  - Number of trees
  - Tree depth
  - Min sample for split
  - Min samples for leaf

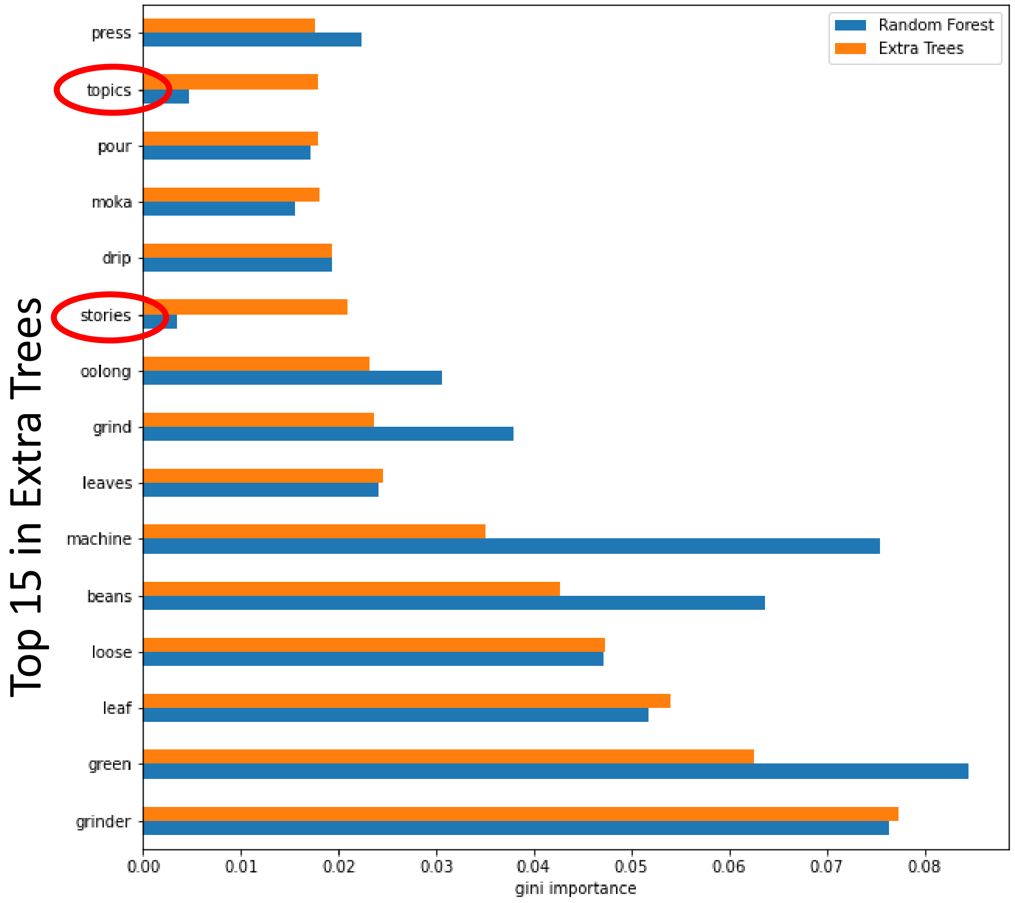
- Which is better?

Dataset	Accuracy scores	
	Random Forest	Extra Trees
Cleaned	0.832	0.812
Lemmatized	0.850	0.833
Stemmed	0.862	0.850

Random Forest vs Extra Trees

- Words in spam posts has higher importance in Extra Trees.
- Not found in Random Forest.

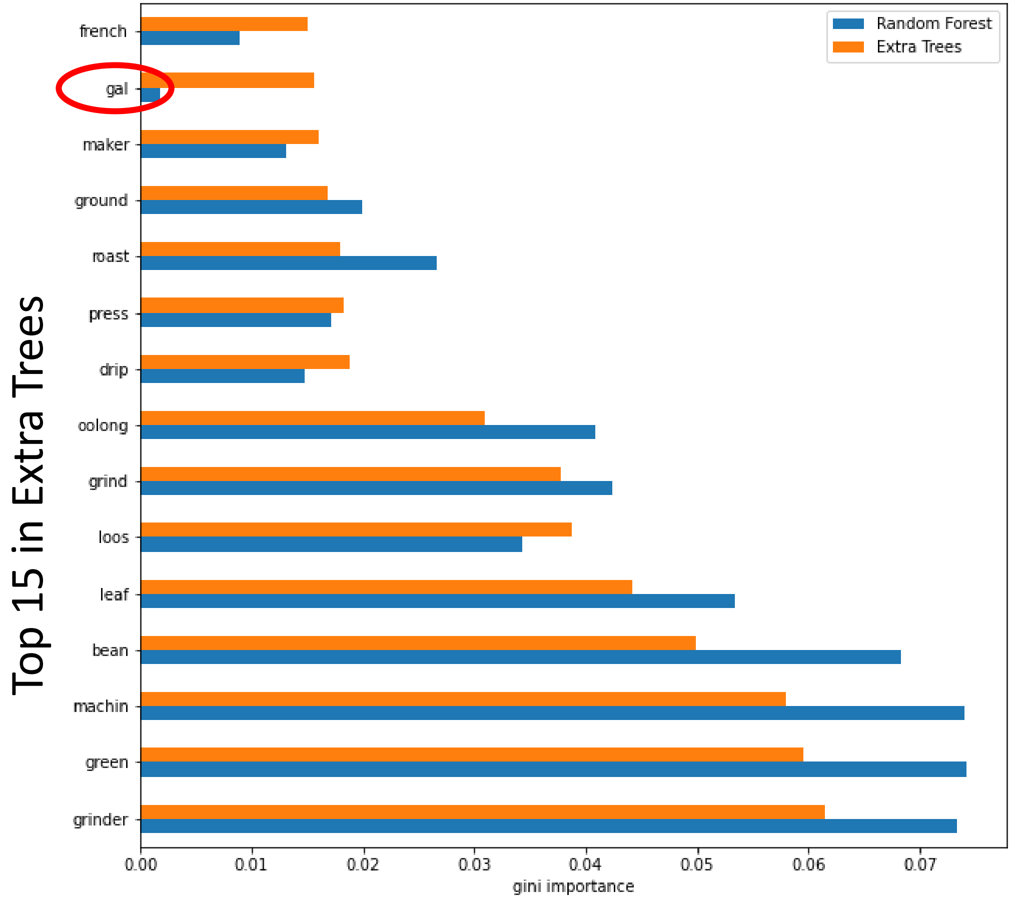
Cleaned dataset



Random Forest vs Extra Trees

- Stemmed features?
- ‘gal’ in spam posts has higher importance in Extra Trees.
- Not found in Random Forest.

Stemmed dataset



Random Forest vs Extra Trees

- Metrics:
- Random Forest has higher accuracy and f1 scores than Extra Trees.

- Both models performed better for stemmed dataset.
- If there were many irrelevant features, stemming might be better at consolidating features.

Conclusions

- Naïve Bayes model the model of choice.
  - It provided good accuracy and f1-scores, compared to other models.
  - It reported good scores for all datasets.
  - The model performed well *despite* the presence of spam.

- Recommendations:
  - Better cleaning
  - Validity longer term

Thanks
Thanks
Thanks
Thanks
Thanks
Thanks
Thanks