AUTOMATED LANGUAGE RECOGNITION TOOL

Ankita Singh & Nimisha Srinivasa
UCSB CS 273 Fall 2016

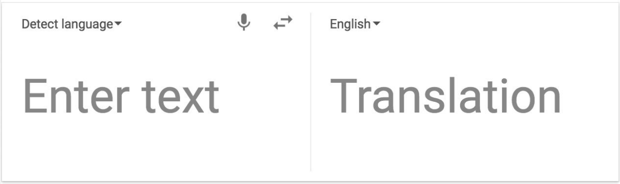
OUTLINE

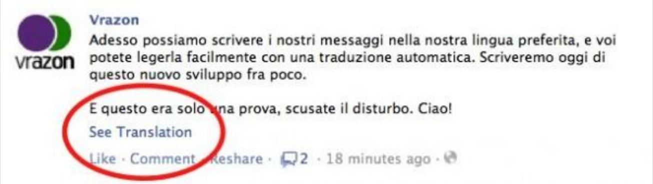
- Problem statement
- Challenges
- Dataset
- Methodology
- Results and analysis
- Takeaways/Thoughts

MOTIVATION

Build an automated **Language Classifier** for Twitter.

- Analysis of **short texts** on social media is gaining importance.
- Language detection is currently focussed on **long written text**.
- A lot of new languages are becoming better represented online.





Adesso possiamo scrivere i nostri messaggi nella nostra lingua preferita, e voi potete legerla facilmente con una traduzione automatica. Scriveremo oggi di questo nuovo sviluppo fra poco.

E questo era solo una prova, scusate il disturbo. Ciao!

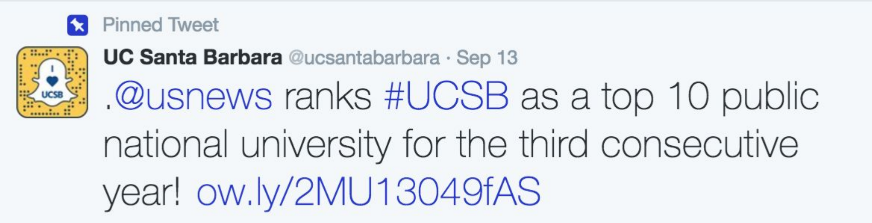
Challenges

- Tweets are only 140 characters long
- Colloquial language and misspellings
- Abundant use of hashtags, handles, URLs and emoticons.

DATASET

Twitter dataset:

https://blog.twitter.com/2015/evaluating-language-identification-performance

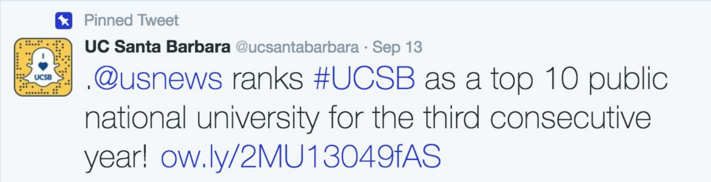


Pinned Tweet

UC Santa Barbara @ucsantabarbara · Sep 13

.@usnews ranks #UCSB as a top 10 public national university for the third consecutive year! ow.ly/2MU13049fAS

DATA PREPROCESSING



Pinned Tweet

UC Santa Barbara @ucsantabarbara · Sep 13

.@usnews ranks #UCSB as a top 10 public national university for the third consecutive year! ow.ly/2MU13049fAS

Removing:

- Emoticons
- Digits & punctuations
- Handles & hashtags
- URLs

```
{
        “Id” : ”123456”,
        “Content”  : ”ranks as top public national university for the third consecutive year”,
        “Label” : ”English”
}
```

LANGUAGES

- Arabic
- Latinized Arabic
- Bulgarian
- Bosnian
- Catalan
- Czech
- Danish
- German
- Greek
- English
- Spanish
- Persian
- Latinized Hindi
- Finnish
- French
- Guliguli
- Hebrew
- Hindi
- Croatian
- Haitian Creole
- Hungarian
- Indonesian
- Italian
- Japanese
- Latinized Japanese
- Javanese
- Khmer
- Korean
- Latinized Korean
- Latvian
- Mongolian
- Marathi
- Malay
- Nepali
- Dutch
- Norwegian
- Polish
- Portuguese
- Romanian
- Russian
- Albanian
- Serbian
- Sudanese
- Swedish
- Swahili
- Tamil
- Thai
- Tagalog
- Turkish
- Ukrainian
- Urdu
- Latinized Urdu
- Vietnamese
- Xhosa
- Simplified Chinese
- Traditional Chinese

FEATURE EXTRACTION

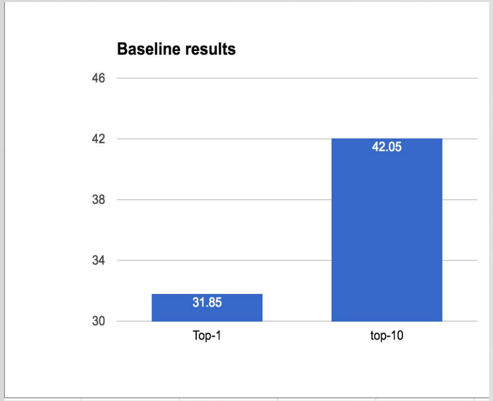
- Frequency features
  - Top-k most frequently used words for every language
- N-gram features: better performance[1]

[1] Grefenstette, Gregory (2014).”C*omparing two language identification schemes*”.

IMPLEMENTATION

BASELINE

- Top-1 most frequent word for each language
- Top-k most frequent words for each language



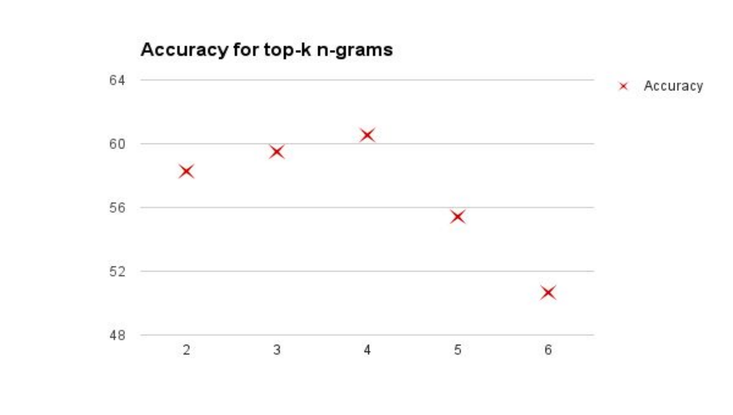
BASELINE

Shortcomings

- Huge dictionary size
- Missing spaces between words for some languages.

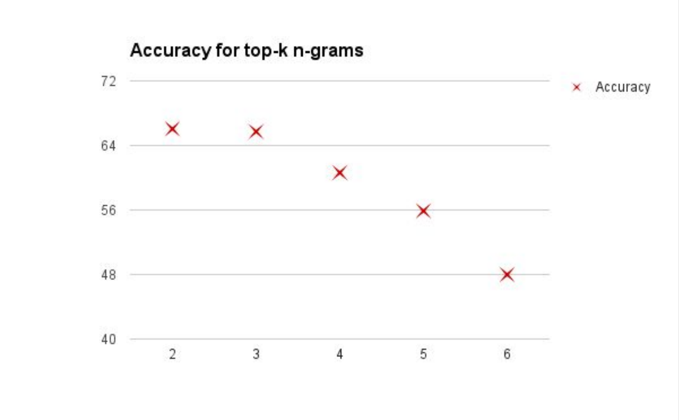
NAIVE BAYES CLASSIFICATION

$$\widehat{L}_k = \underset{L_k}{arg\max}\ p(L_k|x_1, x_2, \cdots x_n) = \underset{L_k}{arg\max} \sum_{i=1}^{n} log(p(x_i|L_k))$$



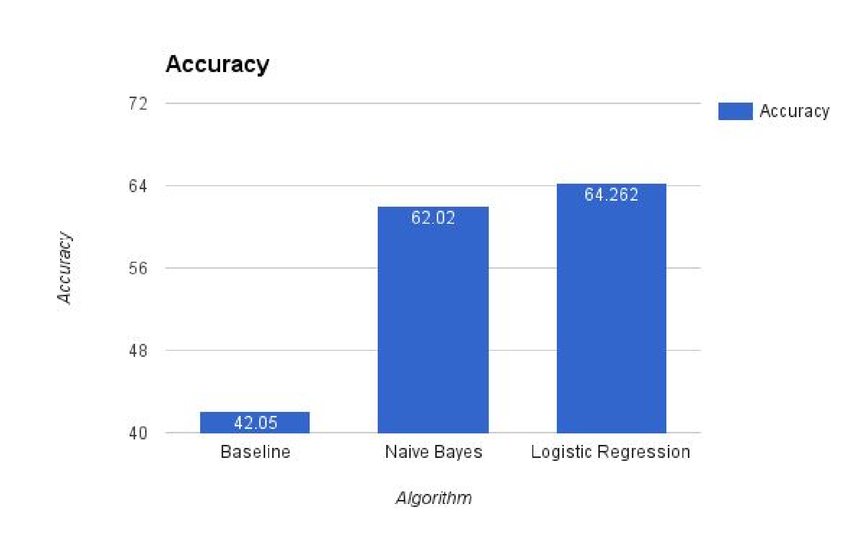
LOGISTIC REGRESSION

$$Pr(Y = c \mid X = x) = \frac{e^{\beta_0{}^{(c)} + x.\beta^{(c)}}}{\sum_c e^{\beta_0{}^{(c)} + x.\beta^{(c)}}}\}$$



Better accuracy compared to NBC!

RESULTS


Accuracy
72
64
56
48
40
Accuracy
42.05
62.02
64.262
Baseline
Naive Bayes
Logistic Regression
Algorithm
Accuracy

LIMITATIONS

- Data set covers just 70 languages.
- Not enough data to process for each language
- Issues in classifying certain very similar languages.

CONCLUSIONS & FUTURE WORK

- Language classifier for tweets
- Best result: Logistic Regression with 3-4 n-grams
- RNN to be considered in future
- Training using the DSL shared task dataset

THANKS!

Code:

https://github.com/nimisha-srinivasa/TweetLanguageClassification