Classic and recent (neural) approaches to automatic text classification

A comparative study with e-mails in the Spanish language

Juan M. Fernández, Nicolás Cavasin, Marcelo Errecalde

Outline

- Introduction
- Automatic text classification approaches
  - BoW + SVM
  - Word2Vec + LSTM
  - BERT
- Research methodology
- Experimental results
- Conclusions

Introduction

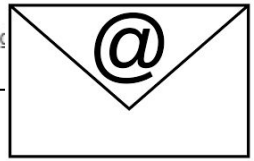
- Text mining became, over the years, one of the most popular fields of research, especially in the field of text classification.
- Diverse approaches have been developed for text representation but most of the resources available are English-centered.

This work presents experiments comparing the performance of the **three most relevant approaches of machine learning applied to text classification** in order to measure how beneficial their contributions to non-English languages are.

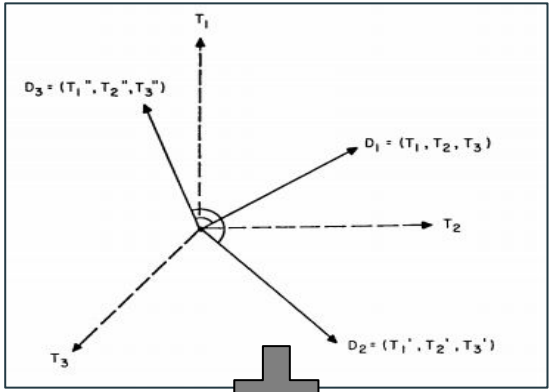
Approach #1: BoW+SVM

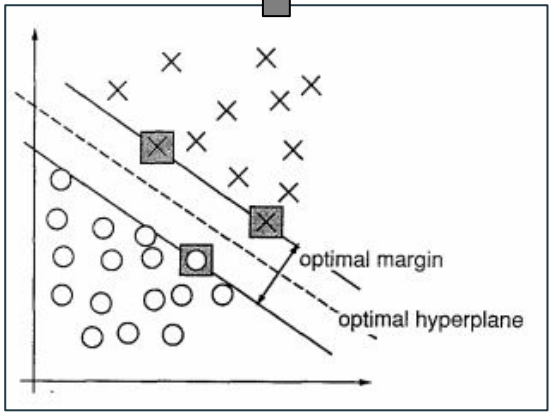
FORMULARIO DE CONTACTO PARA ESTUDIANTES
Enviado :03.24.2021-13:22:08

Nombre y Apellido: Claudia [redacted]
Legajo: 101000
Documento: [redacted]
Carrera: LICENCIATURA EN CS. DE LA EDUCACION(4)
Teléfono: [redacted]
E-Mail: [redacted]
Mensaje / Consulta: Hola queria saber donde teng
dos materias para recibirme.



E-mail preprocessing:
- Text normalization,
- Stopwords removal,
- Static attributes addition,
- Use a variation of n-grams.



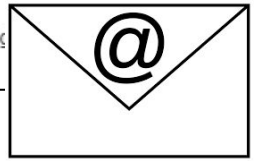


https://github.com/jumafernandez/clasificacion_correos/blob/main/notebooks/jcc/01-bow+binario+svm.ipynb

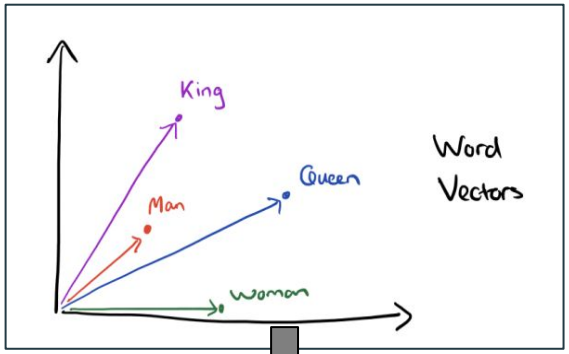
Approach #2: Word2Vec+LSTM

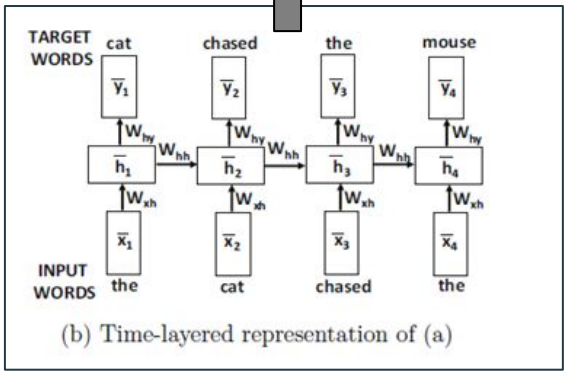
FORMULARIO DE CONTACTO PARA ESTUDIANTES
Enviado :03.24.2021-13:22:08

Nombre y Apellido: Claudia [redacted]
Legajo: 101000
Documento: [redacted]
Carrera: LICENCIATURA EN CS. DE LA EDUCACION(4)
Teléfono: [redacted]
E-Mail: [redacted]
Mensaje / Consulta: Hola queria saber donde teng
dos materias para recibirme.



E-mail preprocessing:
- Text normalization,
- Stopwords removal,
- Pre-trained word embeddings.





(b) Time-layered representation of (a)

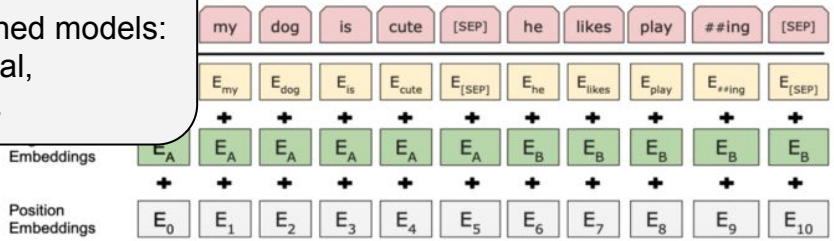
https://github.com/jumafernandez/clasificacion_correos/blob/main/notebooks/jcc/02-Word2Vec+LSTM.ipynb

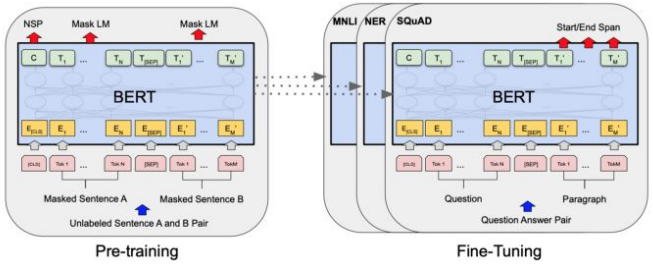
Approach #3: BERT

FORMULARIO DE CONTACTO PARA ESTUDIANTES
Enviado :03.24.2021-13:22:08

Nombre y Apellido: Claudia [redacted]
Legajo: 101000
Documento: [redacted]
Carrera: LICENCIATURA EN CS. DE LA EDUCACION(4)
Teléfono: [redacted]
E-Mail: [redacted]
Mensaje / Consulta: Hola queria saber donde tenç
dos materias para recibirme.

E-mail preprocessing:
- Text normalization,
- Two BERT pre-trained models:
--- Google Multilingual,
--- Spanish (UChile).





https://github.com/jumafernandez/clasificacion_correos/blob/main/notebooks/jcc/03-BERT-wandb.ipynb

Research methodology

- This research work uses a dataset generated from academic questions made by e-mail by students of the National University of Luján to the administrative staff.
- 1000 interactions have been selected and labelled, by a domain expert, on four different classes:
  - Public transport discount ticket,
  - Admission to the university,
  - Admission requirements,
  - Other topics.
- For the experiments, the original e-mails were used without any human supervision on semantic nor syntactic mistakes.

Experimental results

In every approach a search for the best hyper-parameters was applied to each strategy, obtaining the following results:

Strategy	Accuracy	Precision	Recall	F1-score
BoW+SVM	0.870	0.862	0.830	0.840
Word2Vec+LSTM	0.835	0.814	0.841	0.820
BERT (Multilingual)	0.860	0.838	0.842	0.840
BERT (BETO)	**0.890**	**0.870**	**0.885**	**0.876**

Conclusions

Based on the previous results, and **considering the 30 years of evolution** that this discipline has experienced since BOW+SVM first appearance and BERT's presentation, **we have verified that the traditional representation and classification methods are still a very competitive option**.

- However, it is important to keep in mind that e-mails have some features that do not help these cutting-edge models due to its informal manners and syntactic mistakes which are frequently seen in this type of communication model.
- That is why the precision gap between cutting-edge models and traditional ones is expected to maximize when datasets with cleaner texts are used.

¡Thank you!

jmfernandez@unlu.edu.ar
ncavasin@unlu.edu.ar
merreca@unsl.edu.ar

https://github.com/jumafernandez/clasificacion_correos

IX JORNADAS DE CLOUD COMPUTING, BIG DATA & EMERGING TOPICS