Text analysis for company acquisistion

ELENA VELTRONI, FILIPPO MINUTELLA

Purpose of the project

The main goal of the project was to implement an application in which the user can perform clustering and classification analysis on company acquisition articles.

Dataset composition

Data by Faculty of Economy

- Data are extracted from selected articles by different sites
- Articles are divided in sentences and the more interesting sentences are selected
- A sentiment is assigned to the selected sentences

Data by Twitter

- Part of data are retrieved from Twitter
- Free twitter API limit the possibility to retrieve only tweet published in the previous 7 days
- We use GetOldTweets3 library for Python to search specific Twitter in a specific period

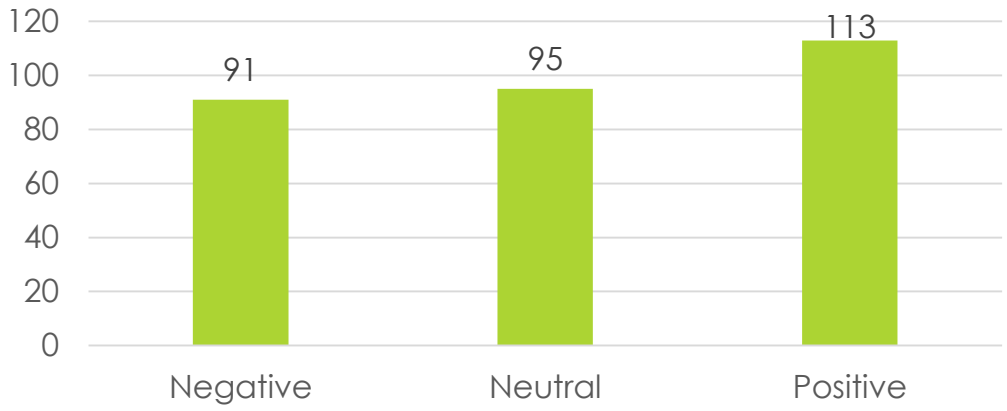
Dataset

Imbalanced

- The first version of the dataset received was highly imbalanced.
- This caused poor efficiency in the first classification models trained.
- The class with lower examples was always misclassified

Balanced

- We requested to balance the dataset
- The final dataset presented 299 entries



Text to Vector

We need to transform text to vector

BoW

- Simpler and faster model
- Language agnostic

BERT (Sentence Transformer)

- More accurate
- Language dependent

BoW

This model works in an unsupervised way

Based on word frequencies

Bag of Words Example

Document 1

The quick brown fox jumped over the lazy dog's back.

Document 2

Now is the time for all good men to come to the aid of their party.

Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

Stopword List

for
is
of
the
to

Sentence Transformer

bert-base-nli-mean-tokens

BERT makes use of Transformer, a mechanism that learns contextual relations between words in a text. Transformer includes two separate mechanisms, an encoder that reads the text input and a decoder that produces a prediction for the task.

Classification with BERT

The user can choose the classification model to use in prediction

In the training phase 3 classifiers are trained and its performance tested with cross validation (10-folds) are printed out:

- KNN
- SVM
- Logistic Regression

Confusion matrices

K-NN

	a	b	c
a = -1	69	9	13
b = 0	41	17	37
c = 1	28	26	62
Precision	0.5	0.3269230769230769	0.5535714285714286
Recall	0.7582417582417582	0.17894736842105263	0.5344827586206896
F-Measure	0.6026200873362445	0.2312925170068027	0.5438596491228069
Accuracy		0.5017241379310345	
Standard Deviation		0.05399669205919909	

SVM

	a	b	c
a = -1	58	18	15
b = 0	26	35	34
c = 1	18	34	64
Precision	0.5686274509803921	0.40229885057471265	0.5663716814159292
Recall	0.6373626373626373	0.3684210526315789	0.5517241379310345
F-Measure	0.6010362694300517	0.3846153846153846	0.5589519650655022
Accuracy		0.5486206896551724	
Standard Deviation		0.08782098138882705	

Logistic Regression

	a	b	c
a = -1	54	17	20
b = 0	20	41	34
c = 1	18	32	66
Precision	0.5869565217391305	0.45555555555555555	0.55
Recall	0.5934065934065934	0.43157894736842106	0.5689655172413793
F-Measure	0.5901639344262295	0.44324324324324327	0.5593220338983051
Accuracy		0.5220689655172414	
Standard Deviation		0.07282096800301242	

Improve the training set

There are not statistical differences between the classifiers

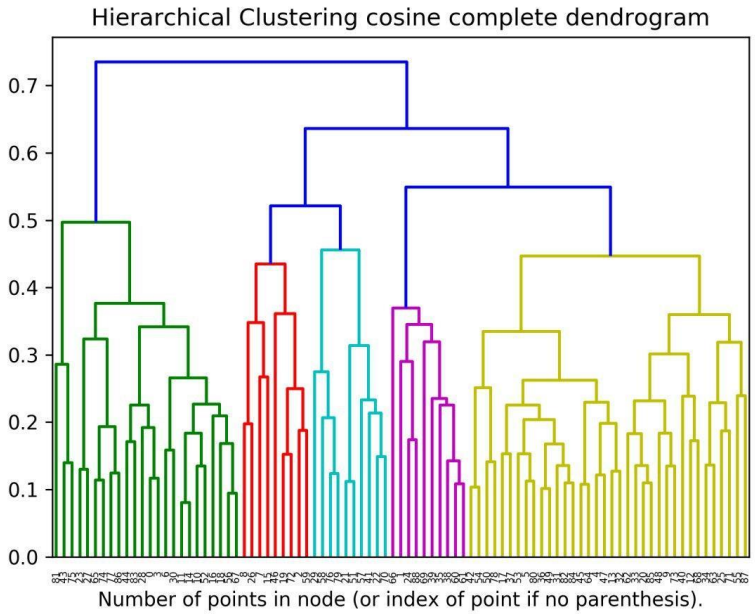
- The web app deployed allows to insert new data to the training set to improve the classifications performances
- Whenever new data are inserted to the training set and the training process is run the new performances info are returned

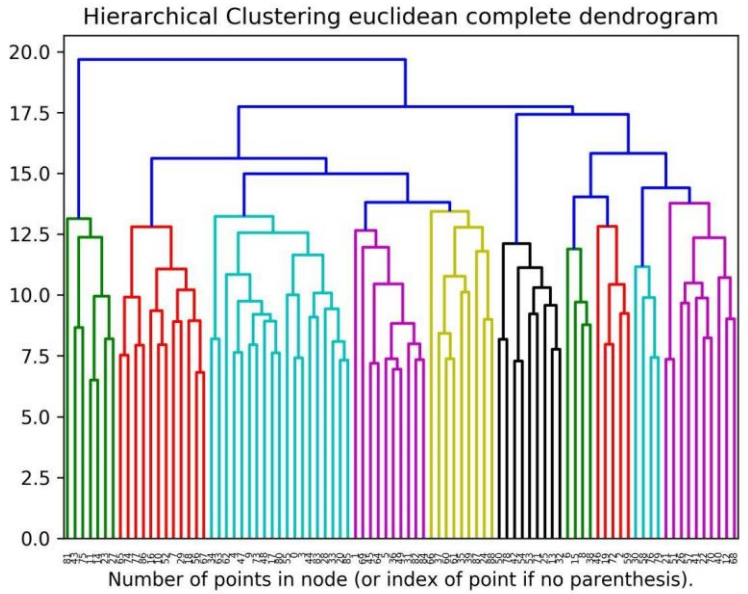
Clustering with BERT

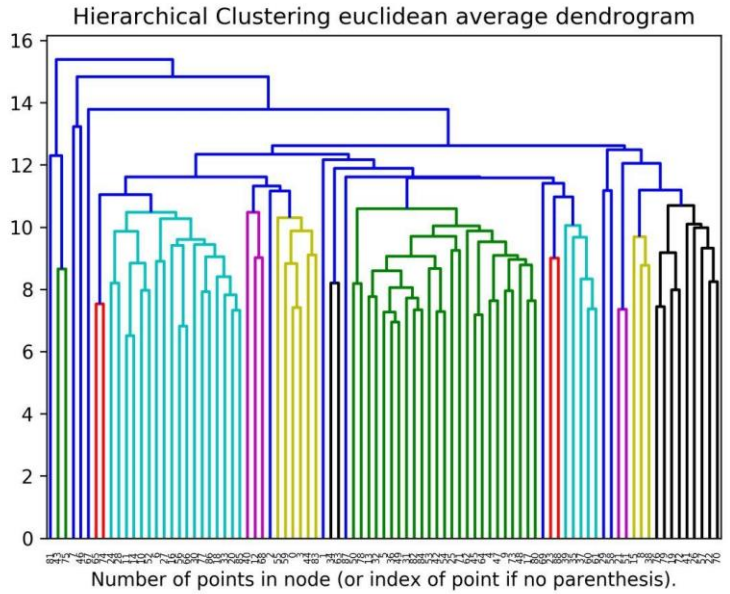
Classification not so good, searching for correlation between articles

- The web app allows to view clustering dendogram choosing linkage metric and affinity measure
- The result is an image with the dendogram and the articles with corresponding cluster

Clustering Dendogram







Application screenshot