



## Data Mining Project

Elena Veltroni, Filippo Minutella

February 2020

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Dataset</b>	<b>4</b>
2.1	Data provided . . . . .	4
2.2	Data by twitter . . . . .	5
<b>3</b>	<b>Text Elaboration</b>	<b>6</b>
3.1	Bag Of Words . . . . .	6
3.2	Bidirectional Encoder Representations from Transformers . . . .	6
<b>4</b>	<b>Classification</b>	<b>9</b>
<b>5</b>	<b>Clustering</b>	<b>11</b>
<b>6</b>	<b>Application</b>	<b>13</b>
6.1	Use Case . . . . .	14
6.2	Application manual . . . . .	15

# **1 Introduction**

The project born with the collaboration with the Faculty of Economy.

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.

In particular, we pre-processed the data provided by the colleagues in the Faculty of Economics, performing a data cleanup in order to prepare data to send to the classifiers and to the clustering algorithm.

The result we obtain can be useful in trading and others analysis performed by Faculty of Economy.

The users can perform the analysis with a simple web app described below.

## 2 Dataset

### 2.1 Data provided

The dataset is composed by some selected sentences in documents that talk about the company acquisitions.

The starting point are articles from the economic sector dealing with company acquisitions. The article is not taken in its totality, but it is divided into sentences, the decision to select a sentence is made by the subjective feeling which the sentence give to the user, only for these sentences a class is assigned.

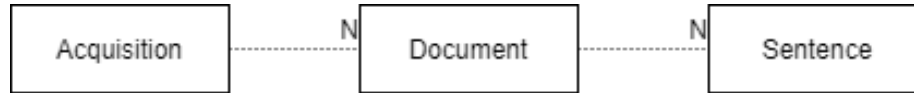


Figure 1: Dataset organization

When a sentence is selected the user classified it with a label indicating if the sentence is *Negative*, *Neutral* or *Positive* according to the user feeling.

This job is done by the colleagues in the Faculty of Economy that provide us an excel file containing the labeled sentences.

In this file we had to perform some cleaning tasks, such as delete sentences written in Italian (the almost totality of the sentences was in English) and make the file interpretable for a Python script.

The first version of the dataset provided to us was highly imbalanced, this caused, in the first version of the classification algorithms, poor classification performances.

So with the objective to improve the classification performances the dataset was balanced and the current class disposition is:

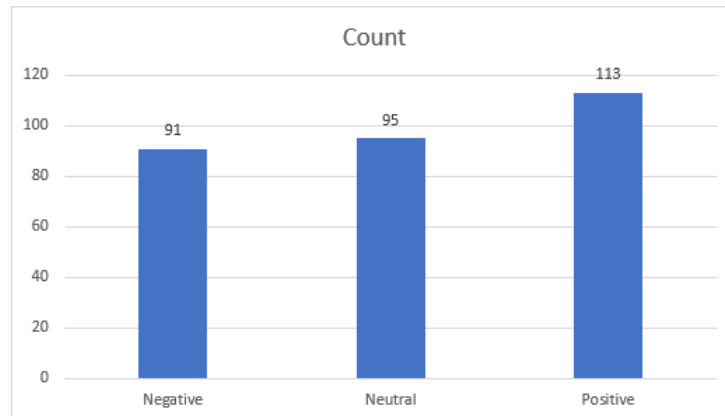


Figure 2: Dataset balanced

## 2.2 Data by twitter

The dataset of this application also contains a part of data retrieved from Twitter.

Twitter makes available to external users official APIs to search for tweets within the platform, restricting the search for tweets by specifying some parameters, but they show a big limit that is the possibility to retrieve only tweets published in the previous 7 days. To get around this limitation, the *GetOldTweets3* library for Python has been used, a library that allows to search by username or keyword, and allows us to specify the period of the tweets wanted.

### 3 Text Elaboration

There not exists algorithm that works with sentences directly so we have to get numbers from sentences that represent the sentences to fit the algorithms.

#### 3.1 Bag Of Words

We have followed two way, the first, the simpler, uses Bag of Word (*BoW*) algorithm.

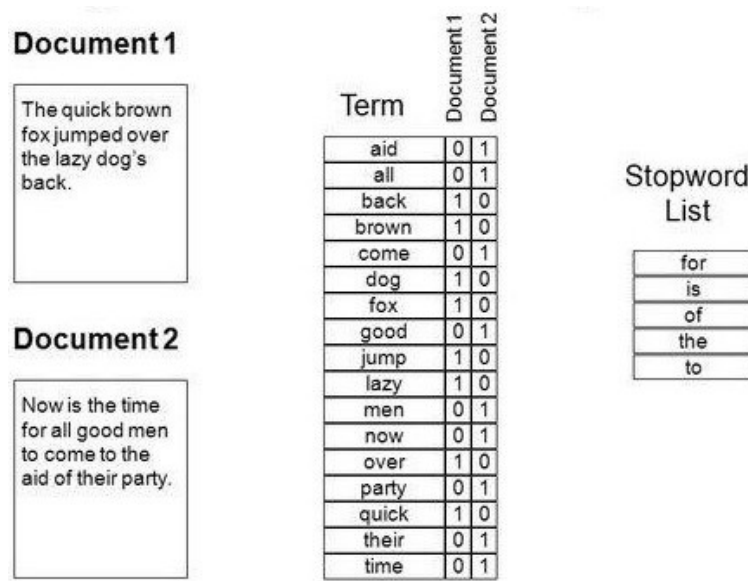


Figure 3: BoW Example

*BoW* works creating vector that represent the presence or absences of a word in a sentence. The words taken under consideration for presences is the K words most frequent in the dataset excluding the stop words.

#### 3.2 Bidirectional Encoder Representations from Transformers

The second way used to elaborate the text is using Sentence-Transformers a repository that fine-tunes *BERT* / *RoBERTa* / *DistilBERT* / *ALBERT* / *XLNet* to produce semantically meaningful sentence embeddings.

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.

As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once, therefore it is considered bidirectional. This characteristic allows the model to learn the context of a word based on all of its surroundings.

BERT uses two training strategies:

- The first model is called Mask Language Model (*MLM*) - and is used to predict a few words and self-check that you have actually understood what you are talking about.

Before feeding word sequences into BERT, 15% of the words in each sequence are replaced with a [MASK] token. The model then attempts to predict the original value of the masked words, based on the context provided by the other, non-masked, words in the sequence.

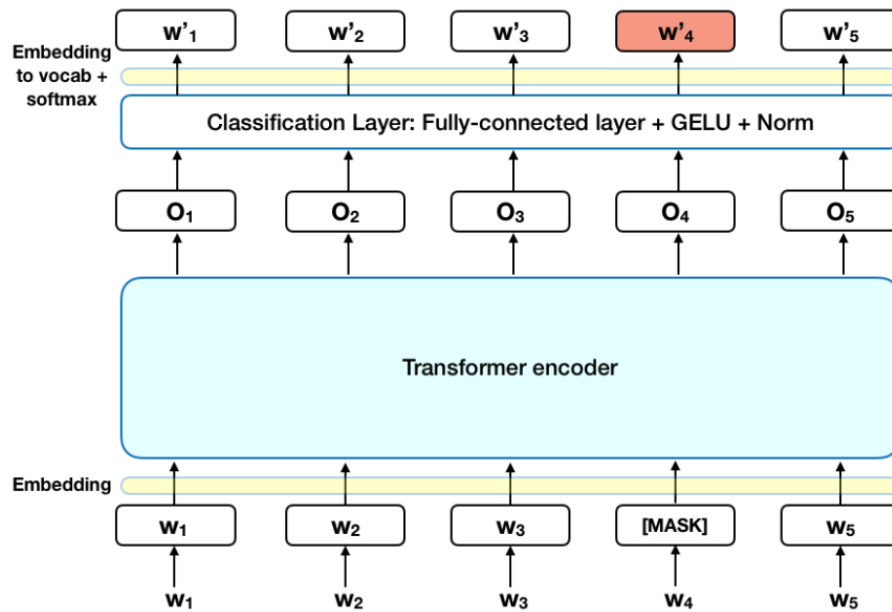


Figure 4: MLM training

- The second technique is called Next Sentence Prediction (*NSP*) and is used by BERT to relate sentences to each other. In the training process, the model receives pairs of sentences as input and learns to predict whether the second sentence of the pair is the next sentence in the original document.

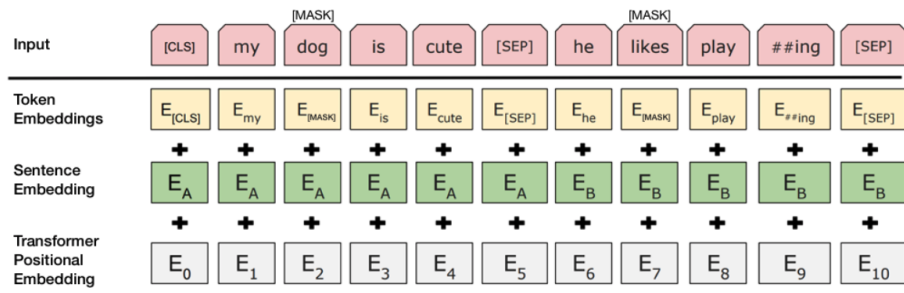


Figure 5: NSP training



## 4 Classification

We perform classification using 3 algorithms: K-NN, SVM, Logistic Regression. These algorithms was trained using the 768 features returned by the Sentence-Transformers (using the model bert-base-nli-mean-tokens) and the class presented in the dataset.

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		

Figure 6: K-NN confusion matrix

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	0.558951965065022
Accuracy	0.5486206896551724		
Standard Deviation	0.08782098138882705		

Figure 7: SVM confusion matrix

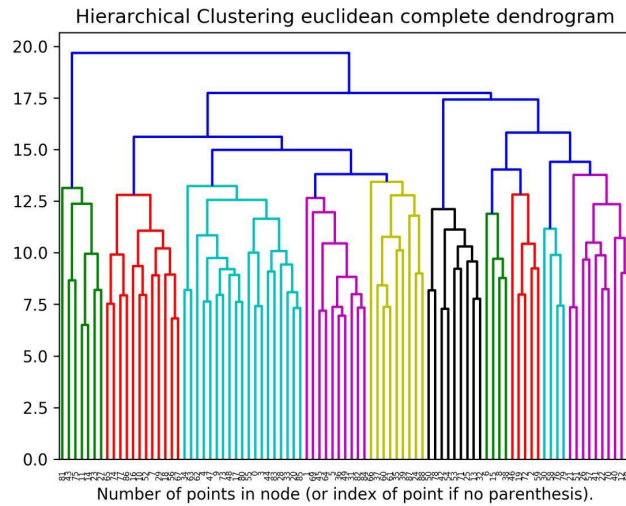
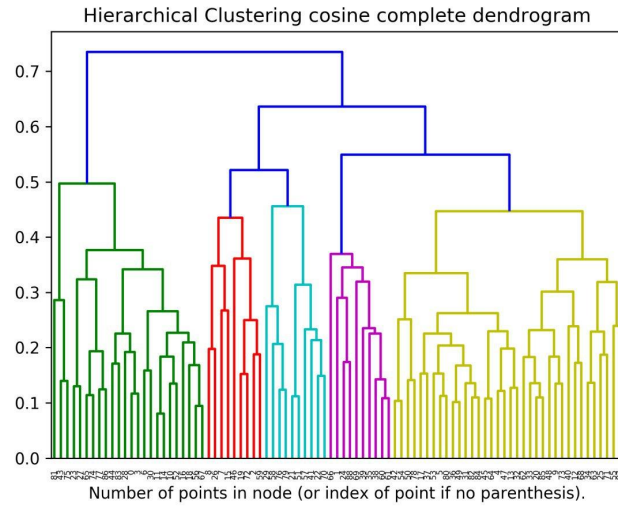
Logistic Regression			
	a	b	c
a = -1	54	17	20
b = 0	20	41	34
c = 1	18	32	66
Precision	0.5869565217391305	0.4555555555555555	0.55
Recall	0.5934065934065934	0.43157894736842106	0.5689655172413793
F-Measure	0.5901639344262295	0.44324324324327	0.5593220338983051
Accuracy	0.5220689655172414		
Standard Deviation	0.07282096800301242		

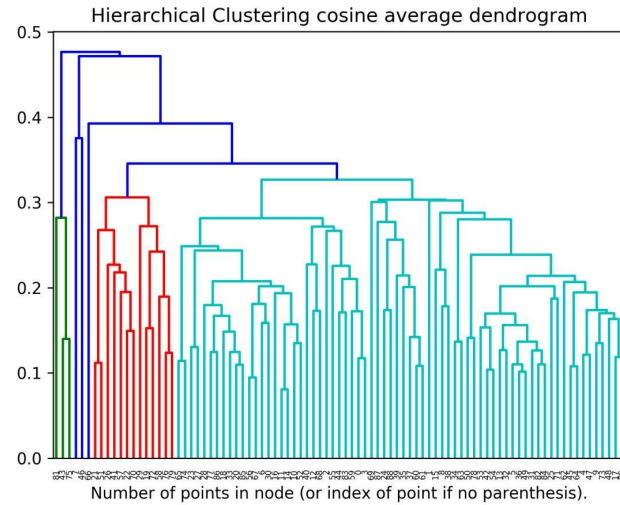
Figure 8: Logistic Regression confusion matrix

The classification performances of the three classifiers are not so good, we have low precision and recall specially for the neutral class.

## 5 Clustering

The clustering was performed using an agglomerative clustering with different linkage metrics and affinity measures based on the BERT returned features. The results of the different clustering are the dendrograms and the cluster membership of the sentences.





Not all dendrograms have a clear organization of the cluster, in the last one is not clear how the sentences are grouped while the first two present a clearer organization so its more simple decide a threshold to cut the dendrogram and get the class memberships. These clustering methods allow us to interpret how the cluster are formed and then provide the cluster membership to colleagues in the Faculty of Economy to evaluate the created cluster.

## 6 Application

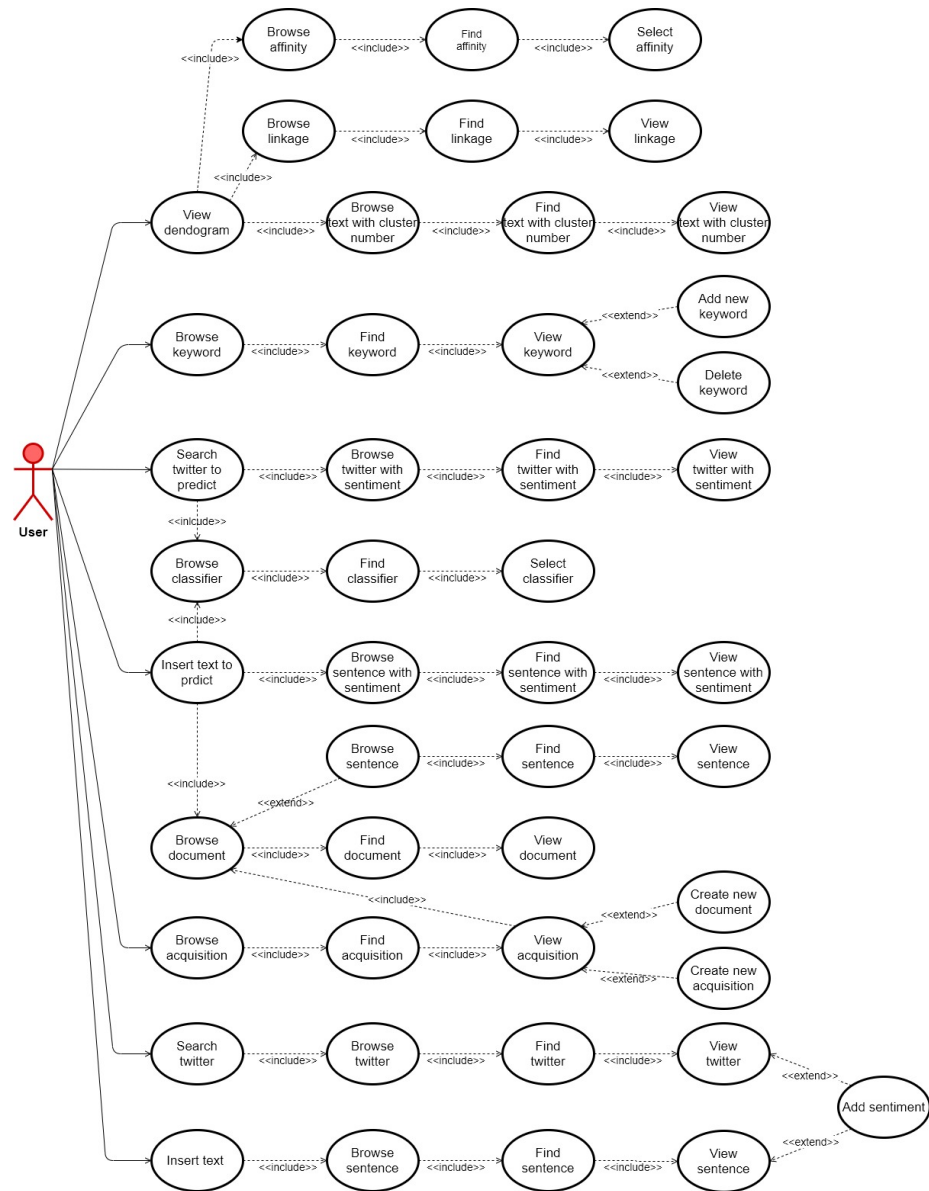
All the functionalities seen before, are implemented in a web app developed using AngularJS for the front-end and Python for the back-end. The application uses MongoDB to store all the data used in the application.

The main actor of the application is a generic user who will use the functionalities offered by the web app.

The main actions allowed by the application are the following:

- Predict the sentiment of a sentences, choosing the classifier.
- Visualize the clustering dendogram, the text and the relative cluster number, choosing the affinity measure and linkage metric.
- Visualize the analytics, in order to see the acquisitions, documents and related sentences stored in the DB.
- Add new acquisitions and documents.
- Manage the keyword in order to improve the accuracy of filtering on the sentences to be predicted.
- Add new data to the training set to improve the accuracy of the classifier.

## 6.1 Use Case



## 6.2 Application manual

Once started the application the window below will be displayed.

With the tab it is possible to choose the type of data on which to make the sentiment prediction, text or twitter.

With the first tab you must enter the entire text and the document to which it belongs. The text will be divided into sentences, splitted or by dot or by a special character specified by the user, and only those containing at least one keyword, specified in the appropriate section, will be shown.

The screenshot shows a web application titled "Sentiment prediction". At the top, there are two tabs: "Insert text" (active) and "Search Twitter". Below the tabs, there is a section "Select a document" with a dropdown menu showing "lexis nexis - analyst transcript". Below this is a large text area containing a paragraph of text: "Deutsche Bank analysts Gregg Gilbert and Greg Fraser said; they believe the combination of the two drug makers makes a best-in-breed company with strong generic and branded franchises. geographic diversity, and very solid management. They reiterated their Buy ratings on both companies. Those price targets are based on each drug maker's standalone prospects." Below the text area is a field "Enter the split character". Below that is a section "Select a classifier" with a dropdown menu showing "SVM". At the bottom right is a blue button labeled "RUN".

Figure 9: Home page - text

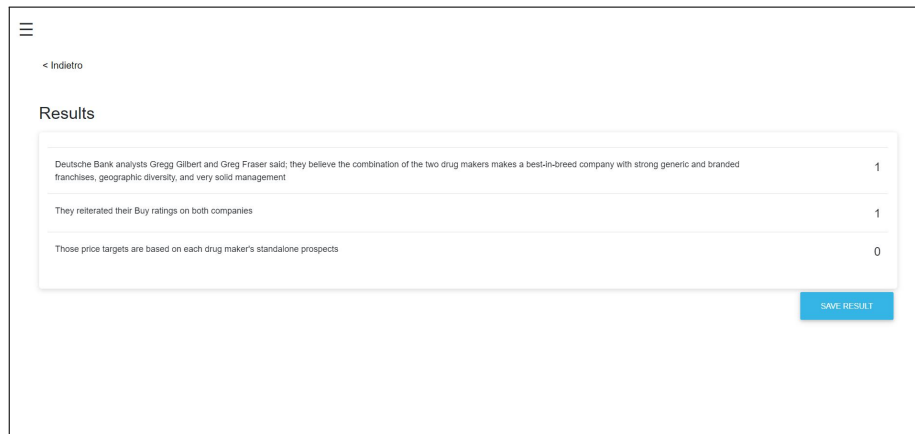
With the twitter tab you can search for keywords or username, specifying @ in front of the word, on twitter by narrowing the search to a specific period.

The screenshot shows the same "Sentiment prediction" application window, but with the "Search Twitter" tab active. The "Insert text" tab is inactive. Below the tabs, there is a section "Enter a username to search on twitter" with a text input field. Below this is a section "Choose your start date" with a text input field showing "gg/mm/aaaa". Below that is a section "Choose your end date" with a text input field showing "gg/mm/aaaa". Below these is a section "Select a classifier" with a dropdown menu. At the bottom right is a blue button labeled "RUN".

Figure 10: Home page - twitter

Before making the prediction, in both cases, it must be specify the type of classifier to use.

Once going forward, the application shows the selected phrases and the related feeling. If the prediction is good, the user has the possibility to add the result to the training set in order to improve the accuracy of the classifiers.

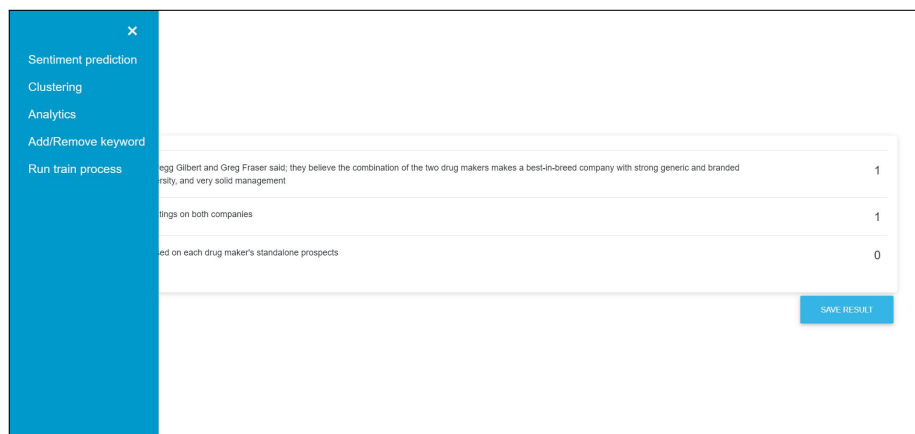


Results	
Deutsche Bank analysts Gregg Gilbert and Greg Fraser said; they believe the combination of the two drug makers makes a best-in-breed company with strong generic and branded franchises, geographic diversity, and very solid management	1
They reiterated their Buy ratings on both companies	1
Those price targets are based on each drug maker's standalone prospects	0

SAVE RESULT

Figure 11: Prediction result

The other features of the application are listed in the side menu.



Results	
Deutsche Bank analysts Gregg Gilbert and Greg Fraser said; they believe the combination of the two drug makers makes a best-in-breed company with strong generic and branded franchises, geographic diversity, and very solid management	1
They reiterated their Buy ratings on both companies	1
Those price targets are based on each drug maker's standalone prospects	0

SAVE RESULT

Figure 12: Menu

The other main function of the application is clustering. On the page below the user can cluster the inserted documents and see how the clusters are formed. The user must choose the linkage metrics and the affinity measure.



Clustering

Enter the threshold

100

Enter the affinity

Euclidean

Enter the linkage

Single Link

VISUALIZE DENDROGRAM

Figure 13: Clustering setting page

After that, the page shows the resulting dendrogram and the list of documents with their cluster number.

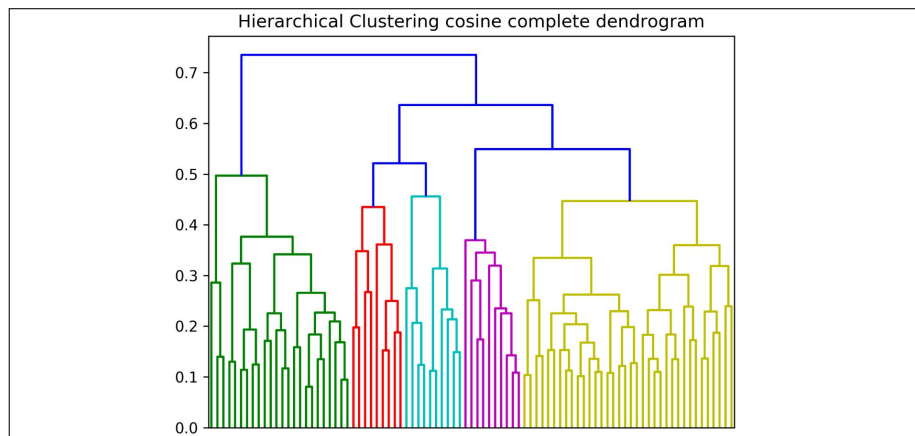


Figure 14: Clustering dendrogram

"Ultimately, this merger is yet another example of the poor incentives Wall Street's quarterly-result mentality creates," Turner added. "Charter would rather take on an enormous amount of debt to pay a premium for Time Warner Cable than build fiber infrastructure, improve service for its existing customers or bring competition into new communities." A heavily indebted Charter Communications will not own the combined entity free and clear. At the close of the deal, Time Warner Cable shareholders will own up to 44% of the new company, Liberty Broadband up to 20%, Advance/Newhouse (Bright House) up to 14%. Charter itself will own just 22%, but will be able to leverage voting control over the entity with the help of Malone's Liberty, which will get almost 25% of the voting power. But the presence of Malone in this deal, even peripherally, is a major concern. Malone-run cable companies are notorious for massive rate increases and poor customer service.	0
We remain bullish on AVGO and expect the stock to sustain its outperformance driven by multiple self-help levers	0
Medco's shares rose 3.3 percent after the companies disclosed the new timelines in securities filings, as investors bet that the risk that U.S. antitrust regulators would block the deal was disappearing in separate filings with the U.S. Securities and Exchange Commission. Express Scripts and Medco said they expected that they may be in a position to close the transaction as soon as the week of April 2. Previously, the companies had said they expected the deal would be completed by the earlier part of the second quarter. Analysts have speculated that the companies may need to sell off one of their specialty pharmacies or mail-order facilities to satisfy FTC concerns. Large grocery chains, many of which operate their own pharmacies, community pharmacies and some consumer groups, have come out against the deal, saying a combined Express-Medco would gain too much power in the market and would squeeze them financially	0
As previously announced, the company expects synergies of \$1 billion once fully integrated, which represents approximately 1 percent of the combined company's costs. The transaction is expected to be slightly accretive to earnings per share (excluding integration and deal-related costs and charges) in the first full year after closing and moderately accretive once fully integrated.	0
he primary driver of our proposal is to increase Boston Scientific's diversification and grow our cardiac-rhythm management business," Boston Scientific's chief operating officer, Paul A. LaViolette, said in an interview with the Associated Press. Boston Scientific executives said the prospect of entering the lucrative \$10 billion market for implantable pacemakers and defibrillators by purchasing Guidant outweighs the legal risks posed by Guidant's recent problems.	0
Upon close, anticipated in the second quarter of next year each delivering strong revenue and earnings growth. Our ability to create sustainable value will be based on delivering double-digit revenue and earnings growth; a diverse revenue base; world-class blockbuster franchises with such well-known names as Botor; and unparalleled global commercial reach. And we are committed to investing in an industry-leading pipeline that will continue to deliver long-term growth. Looking at the chart you can see that our branded pharmaceutical business has a 10% growth CAGR, putting us among the elite growth companies. We also forecast double-digit accretion within the first 12 months, doubling in the next year. This is truly impressive performance, and we anticipate free cash flow in 2016 of more than \$8 billion. We will refine these estimates as we move forward, but we anticipate financial synergies of approximately \$500 million; R&D synergies of approximately \$400 million; COGS savings of \$150 million to \$200 million; sales and marketing savings of approximately \$400 million; and G&A savings in the range of \$350 million. We are in a unique position as a combined company. We will have six blockbuster therapeutic categories with strong well-known global brands we also add a powerful new therapeutic	0

Figure 15: Clustering result

The analytics page summarizes all the acquisitions, documents and sentences with the respective sentiment saved in the DB.

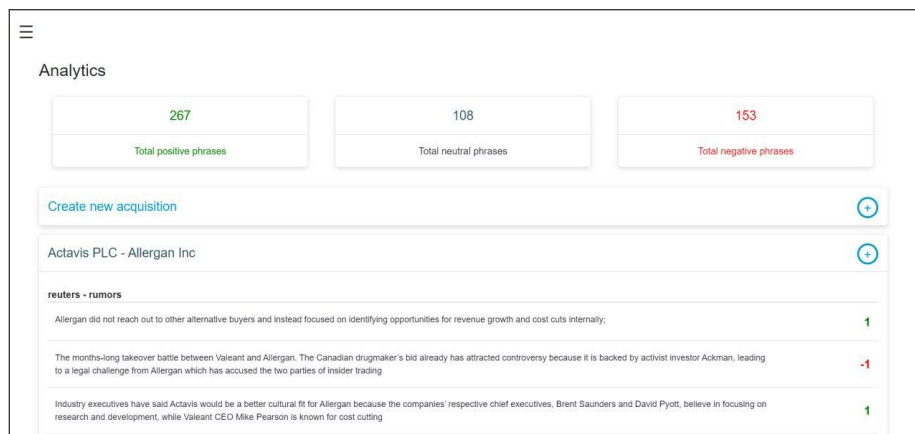


Figure 16: Analytics page

In this page the user also has the possibility to create a new acquisition or add a new document in one of the already existing acquisitions.

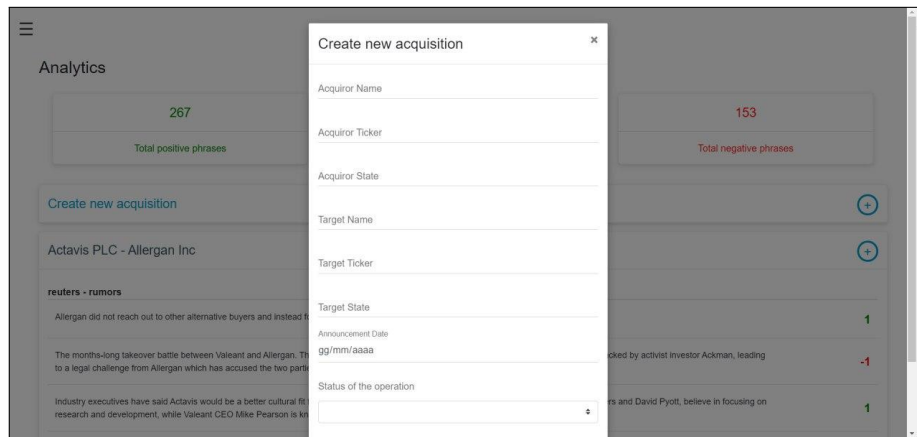


Figure 17: Add acquisition page

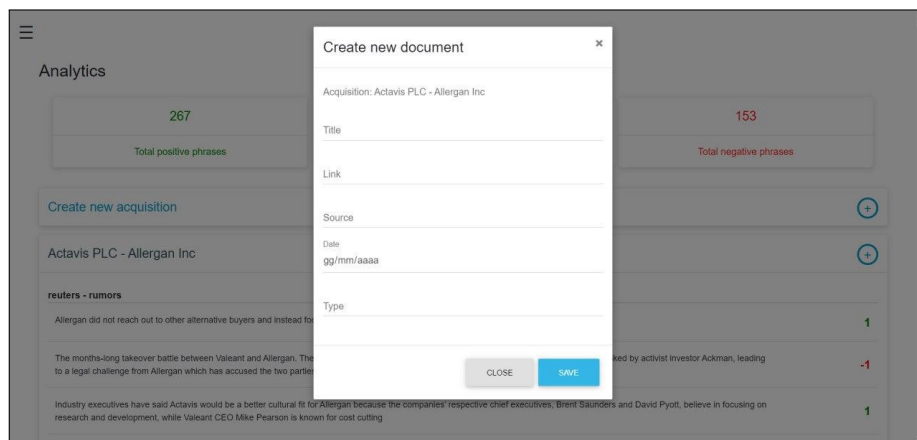


Figure 18: Add document page

The keywords section allows the user to manage them, add new ones or delete those already entered.

Keyword list

[Add new keyword](#)

ebitda

ADD

buyers	X
revenue	X
growth	X
cost	X

Figure 19: Add keyword

The last functionality allowed by the application is the possibility to manually add sentences and the related sentiment to the training set, in order to improve the performance of the classifiers in future predictions. As for the prediction, you can insert phrases of a text or twitter.

Improve the training set

[Insert text](#) [Search Twitter](#)

Select a document

cnbc - news

any posted a better than expected adjusted quarterly profit. The drugmaker also announced that it will adopt the name Allergan. declined to say what percentage of revenue the company would allocate to research and development, saying that metric is "horrible" and has gotten big pharma in a lot of trouble. The company's net loss widened to \$732.9 million, or \$2.76 per share, in the fourth quarter that ended Dec. 31 from \$148.4 million, or 86 cents per share, a year earlier. However, on an adjusted basis the company earned \$3.91 per share. Analysts on average had expected \$3.67, according to Thomson Reuters. Net revenue rose 46 percent to \$4.06 billion. Total North American brands revenue nearly tripled to \$1.83 billion.

Enter the split character

RUN

Figure 20: Add new data to training set

Once you have entered the text or made the search of the interested tweets, the page will show a list of sentences to which it is possible to assign a feeling.

< Indietro

### Results

Company posted a better than expected adjusted quarterly profit

1

The drugmaker also announced that it will adopt the name Allergan

1

declined to say what percentage of revenue the company would allocate to research and development, saying that metric is "horrible" and has gotten big pharma in a lot of trouble

-1

The company's net loss widened to \$732.9 million, or \$2.76 per share, in the fourth quarter that ended Dec 31 from \$148.4 million, or 86 cents per share, a year earlier

However, on an adjusted basis the company earned \$3.91 per share; analysts on average had expected \$3.67, according to Thomson Reuters; net revenue rose 46 percent to \$4.06 billion; total North American brands revenue nearly tripled to \$1.83 billion

ADD TO TRAINING SET

Figure 21: Add sentiment

Once the results are saved, the confusion matrices and related statistical measurements are shown to measure the quality of the classifier.

< Indietro

### Results

K-NN

	a	b	c
a = -1	70	8	13
b = 0	42	15	38
c = 1	28	29	62
Precision	0.5	0.28846153846153844	0.5486725663716814
Recall	0.7692307692307693	0.15789473684210525	0.5210084033613446
F-Measure	0.6060606060606061	0.2040816326530612	0.5344827586206897
Accuracy	0.5017241379310345		

Figure 22: The train result