DS4A. Team 62

Public perception of the Colombian compensation funds

An artificial intelligence approach

Andrés Aranguren MSc. student data Science Universitá degli studi di Padova

Summary:

The superintendency of family subsidy is a Colombian entity in charge of supervising the family subsidy system, specifically the family compensation funds, which offers different services of education, recreation, health, tourism and subsidies among others. The superindency wants to know and understand the users feeling and perception of each fund, in order to improve the customer service and general perception.

To solve this, we took the information from Twitter of the funds, due to is a very important network and the information is public. The model extract all the Tweets related to the funds each 6 hours and save them in a database, where are processed each 5 minutes in groups of 500 tweets.

We create a LSTM recursive neural network model customized and trained in our language (the model was trained with a manual classification to identify local language), that classifies each Tweet into positive or negative feeling, in addition, the model classify the service of which each tweet refers grouped in health, subsides, general, culture and others.

The information is displayed in a dashboard, in which in real time and dynamically the user can filter by compensation fund, service and dates, and see the sentiment analysis, the word cloud with the most common words and the amount of tweets per day and hour.

With this information, the superintendency can identify opportunities of improvement of the services, focused in the funds and services with lower feelings. They can also measure the impact of communications or campaigns.

Introduction:

The compensation funds offer a variety of services and benefits, and the superintendency want to measure and identify the customers feelings and perceptions towards them in social networks. With that input, the superintendency can create different reports and focus the marketing campaigns to improve their image and services.

According to the regulations, these funds receive 4% of the social security contributions paid by employers. The objective of these resources it's to improve the quality of life of the families of Colombian workers and give them wellness, through the management and delivery of subsidies and diverse services. Knowing the citizen´s perception of those subsidies and services is pretty important in the sense that they are the owners of those resources, and they have the right to make sure that the money is properly managed, and receive the services offered.

Application Overview:

Through the use of statistical and computational techniques of Sentiment Analysis, and based on information from social networks, especifically from Twitter, we would measure the users perception regarding the Family Compensation Funds. The analysis includes the massive classification of text documents based on the positive or negative connotation of the language used in it, and the segmentation of each one in the different services provided by them.
The user can see in real time (and in time series) the sentiment of the Tweets of the different compensation funds and of the different services in a dashboard.

Data Engineering:

○ **Interactive Front-end**:

The interactive front-end that we created was done with React, that was designed for "building large scale apps with data that changes over time" makes it painless to create interactive user interfaces. React also helps us to design simple views for each state and update it when the data changes.

React has gained a huge adoption over the last several years. The list of websites using React is constantly growing, some of them are AirBnB, Coursera, Dropbox, Expedia, Facebook, HipChat, Instagram, Netflix, Reddit, Salesforce, Twitter, Uber, WhatsApp, WordPress, Yahoo, Zendesk and many others rely on React.

Our front-end is done with React linked with Dash and Streamlit.

○ **Database**:

To solve this problem, we would use JSON structured data to make different data sets. These data will allow us to analyze the information as a whole or in detail for each Family Compensation Found from different dimensions, including timing and analysis of sentiments.

The data set is made up of the Twitter posts, the basic information of the publisher, the date/time, the classification of the post; if its a response, to whom it is responding, mentions, hashtag, number of favorites and retweets. With this information we have the most used words and the number of times it appears, the polarity and subjectivity.

Data is collected from the Twitter social network using Python the 'tweepy' and ‘NLTK’ libraries that use Tweeter's own APIs to collect the information. Twitter allows us to collect retrospective data for up to a week, so we plan to execute the information collection process several times during the project.
In addition to the public data of the social networks, we will obtain information from the Family Compensation Found with reference to the number of users, service, activities, which will allow us to have a standardized measurement.

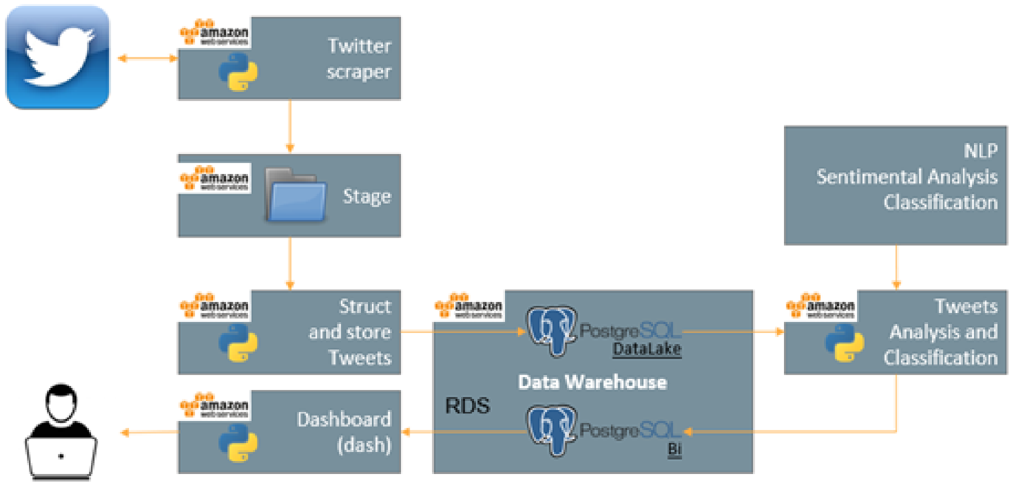
The information is saved in databases in AWS in Postgres.

Code/program design paradigms used:

A Layered Architecture, is the organization of our project structure into four main categories: presentation, application, domain, and infrastructure. Each of the layers contains objects related to the particular concern it represents.
The presentation layer contains all of the classes responsible for presenting the UI to the end-user or sending the response back to the client (in case we’re operating deep in the back-end).
The application layer contains all the logic that is required by the application to meet its functional requirements and, at the same time, is not a part of the domain rules. In most systems that I've worked with, the application layer consisted of services orchestrating the domain objects to fulfill a use case scenario.
The domain layer represents the underlying domain, mostly consisting of domain entities and, in some cases, services. Business rules, like invariants and algorithms, should all stay in this layer.
The infrastructure layer contains all the classes responsible for doing the technical stuff, like persisting the data in the database, like DAOs, repositories, or whatever else you’re using.

Flow charts/diagrams indicating how the different parts interact with each

APP ARCHITECTURE



Data Analysis & Computation:

A fundamental point regarding the project deals with the data visualization and proper results presentation of the tweets traffic and analysis during a specified period. Some of the visualization techniques that will be used are:

- Twitter traffic distribution per week (Identifying peaks and bases), initially for total number of tweets and further graph per each considered compensation fund.
- Map indicating the presence of each fund in the departments of Colombia, and the number of tweets per each fund
- Initial sentiment analysis (polarity) per fund:
    - Gauge Graph: shows the trend or polarity of the sentiment expressed.
    - Wordcloud: represents how often words are used in tweets.
    - Timeline: Represents the trend of sentiment over time.

Models:

- Sentiment analysis: descriptive model including, time series analysis of polarity distribution (positive and negative) in time. Sentiment trend in order to identify possible factors compromising or benefiting funds service quality. Public's sentiment response to major trends and interaction network identifying nodes with the highest traffic (meaning: number of replies per tweet and forwarding frequency).
- Supervised model or transfer learning for tweets classification (requires labelled data): categories will correspond to the funds services (health, insurance, etc),

convolutional neural networks (CNN) and recursive neural network (RNN) may be used using softmax function to classify tweets category.

Naive Bayes Classifier

Sentiment analysis will also be implemented by changing the default parameter of the function textblob.sentiment to a Naive-Bayes analyzer based on the NLTK training algorithm retrieved from reviews database. The main difference from the patternAnalyser, is that the text documents are considered as a bag of words, which is an unordered set of words with their position ignored, keeping their frequency. The Naive Bayes consists basically on a probabilistic classifier, meaning that for a document *d,* out of all classes given $c \in C$,the classifier returns the class c which has the maximum posterior probability in the document.
The basic equation:

$$\hat{c}=argmax\ P(c|d)$$

By substituting the basic conditional probability the equation results:

$$\hat{c} = argmax\ \frac{P(d|c)P(c)}{P(d)}$$

If we want to further develop the equation we can drop the denominator *P(d),* this is possible since we compute the conditional probability for each possible class, hence P(d) doesn't change for each class, we are always asking for the most likely class of the same document *d*.

Datasets + Data Wrangling & Cleaning:

Text preprocessing:

- Stopwords
- Alphanumeric data
- # @ urls

Exploratory Data Analysis:

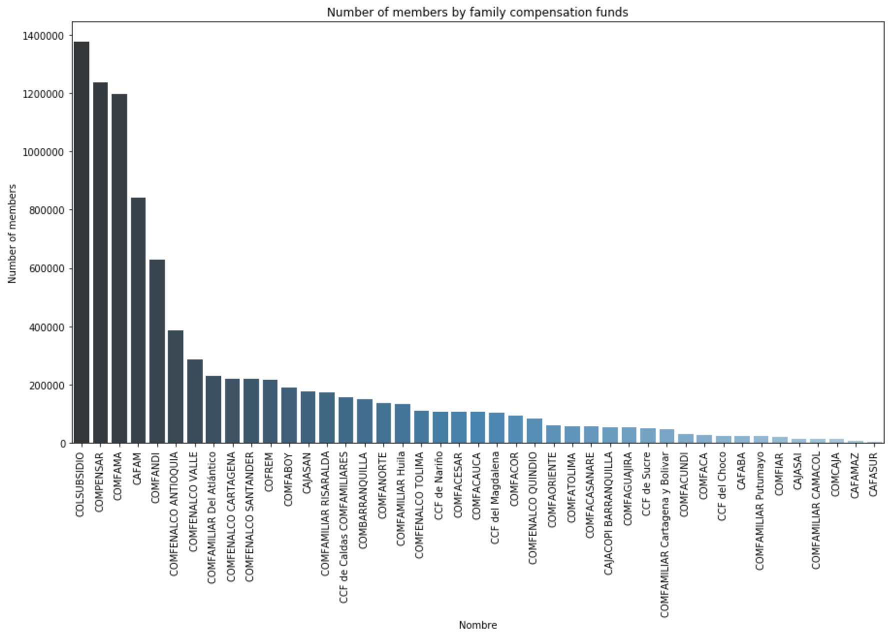


Figure 1. Number of members by family compensation funds.

The 80% of the total members (affiliates) are represented in 15 family compensation funds. Among the top 5 are COLSUBSIDIO, COMPENSAR, COMFAMA, CAFAM and COMFANDI which represents approximately 57% of the total members.

1.2 Amount of tweets per compensation fund:

Compensar is the compensation fund with more mentions in the tweets, representing 55% of the total, followed by Colsubsidio with 18% and Cafam with 12% (Figure 2). It's important to remark that Compensar isn't fund with more affiliates, however is most mentioned suggesting a more higher level regarding digital and online service compared to Colsubsidio (fund with highest number of affiliates)

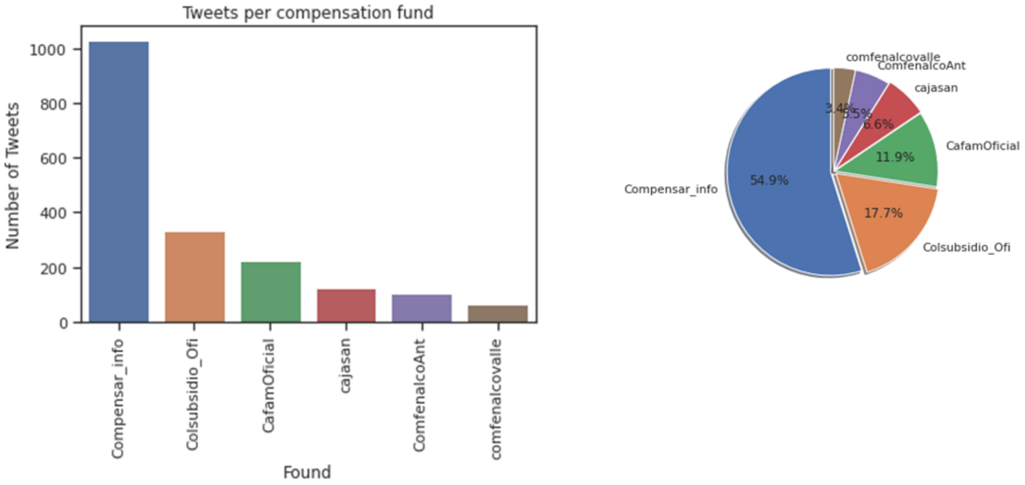


Figure 2. Number of tweets retrieved for main compensation funds

1.3 Tweets traffic daily and weekly distribution

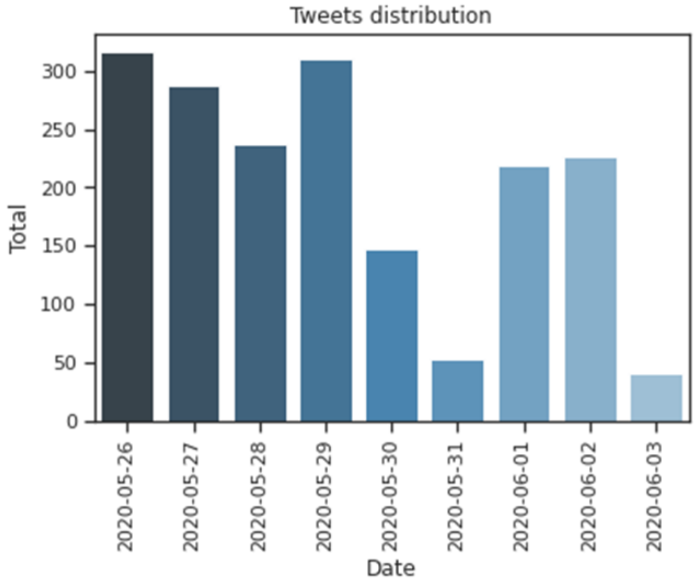


Figure 3. Tweets traffic over retrieval period

The days of the week with more interactions vary in each compensation fund, for Compensar and Cajasan the day with more tweets is on Fridays while for Cafam and Colsubsidio is on Sunday.

The hours with more Tweets is more constant across the compensation funds, the peak hours are in the afternoon between 16:00 and 20:00 range.

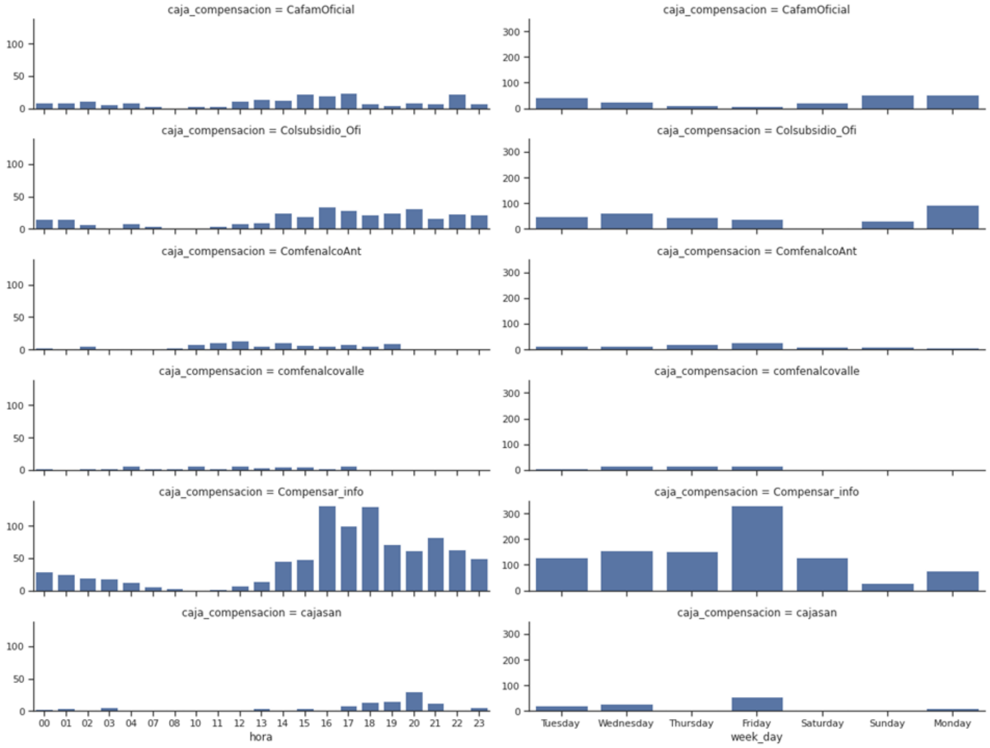
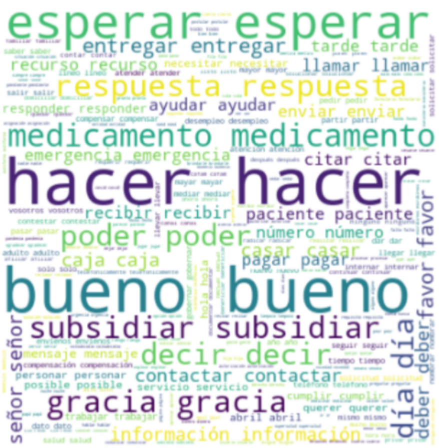


Figure 4. Tweets daily and weekly traffic distribution for main compensation funds

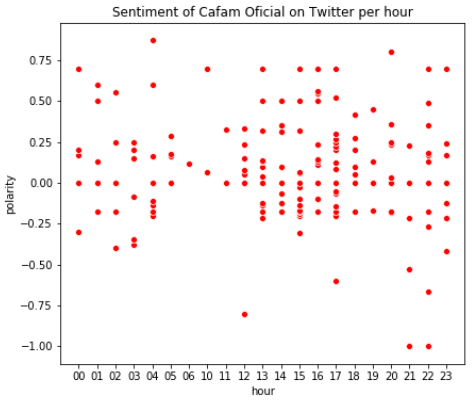
Most common words:

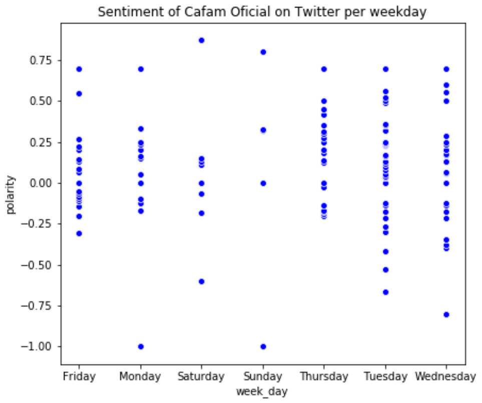
The word cloud of the most common words in the Tweets of the compensation funds are:

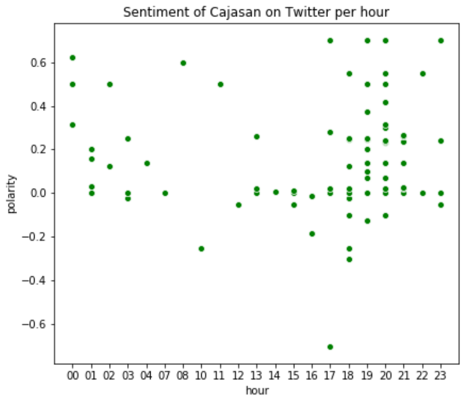


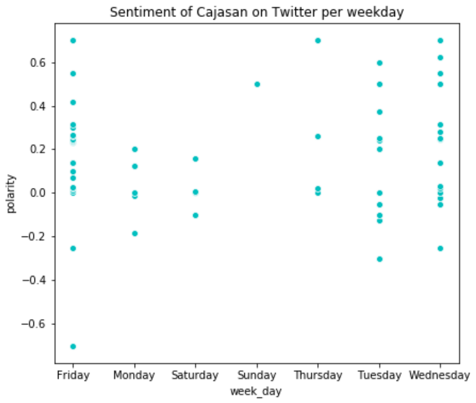
Polarity Sense Analysis

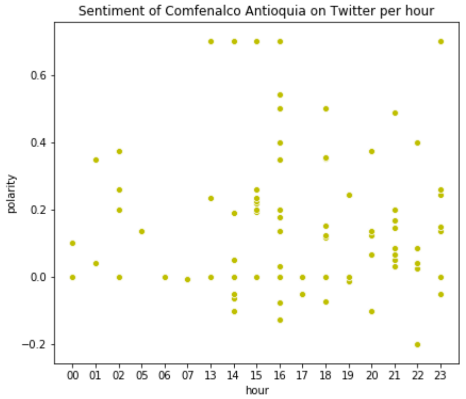
Polarity index values vary between [-1] and [1], where [-1] is a very negative feeling, [0] is a neutral feeling and [1] is a very positive feeling. The analysis of this index disaggregated by compensation fund, hour in the day and day of the week, looks as follows:

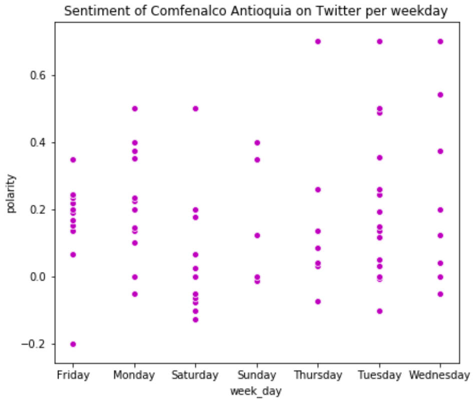

Sentiment of Cafam Oficial on Twitter per hour
polarity
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
00 01 02 03 04 05 06 10 11 12 13 14 15 16 17 18 19 20 21 22 23
hour


Sentiment of Cafam Oficial on Twitter per weekday
polarity
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Friday Monday Saturday Sunday Thursday Tuesday Wednesday
week_day


Sentiment of Cajasan on Twitter per hour
polarity
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
00 01 02 03 04 07 08 10 11 12 13 14 15 16 17 18 19 20 21 22 23
hour


Sentiment of Cajasan on Twitter per weekday
polarity
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
Friday Monday Saturday Sunday Thursday Tuesday Wednesday
week_day


Sentiment of Comfenalco Antioquia on Twitter per hour
polarity
0.6
0.4
0.2
0.0
−0.2
00 01 02 05 06 07 13 14 15 16 17 18 19 20 21 22 23
hour


Sentiment of Comfenalco Antioquia on Twitter per weekday
polarity
0.6
0.4
0.2
0.0
−0.2
Friday Monday Saturday Sunday Thursday Tuesday Wednesday
week_day

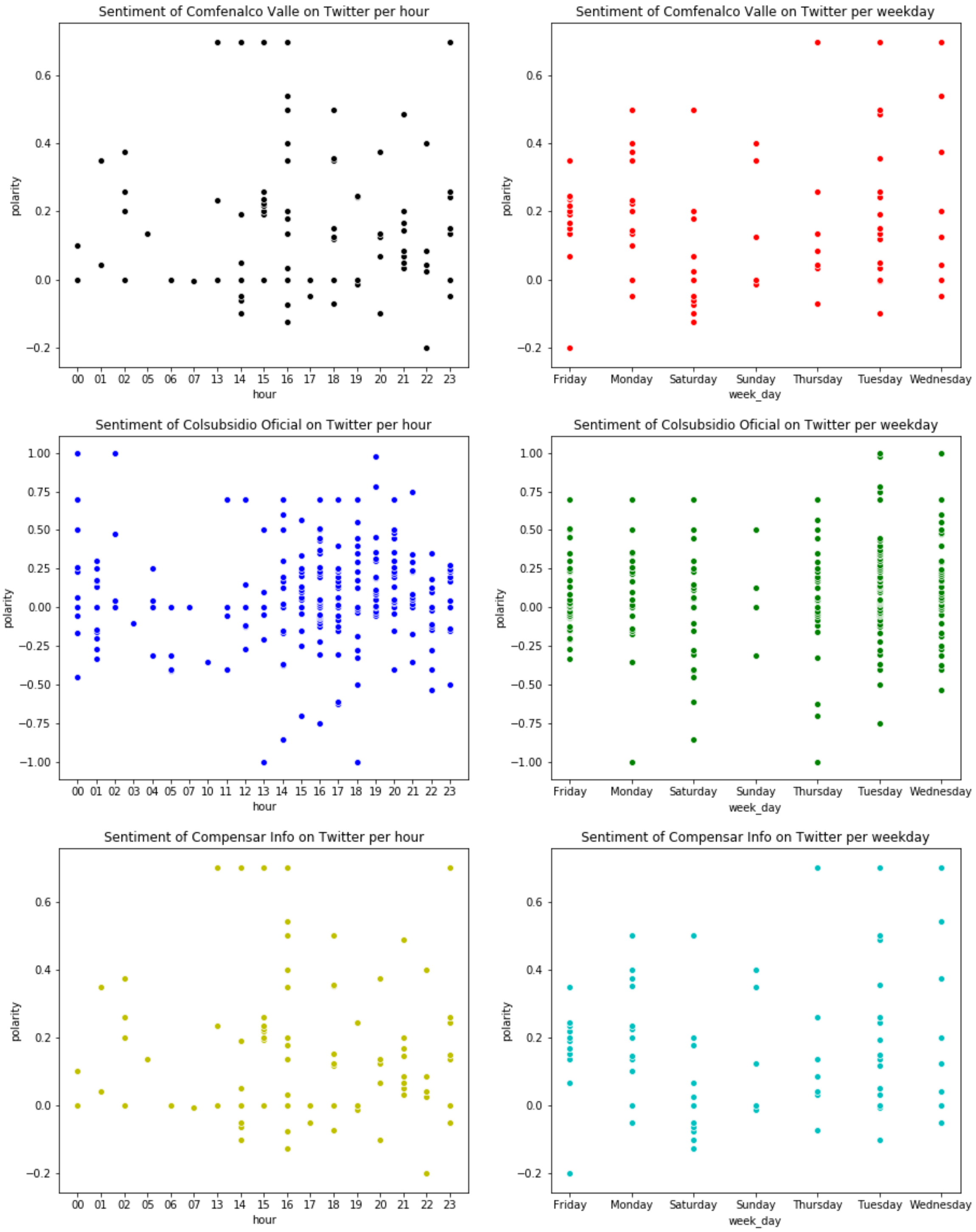


Figure 5.1 Polarity Sense distribution for main compensation funds per hour/per weekday.

Aggregate analysis of polarity can be shown using a Gauge diagram, which is determined from an average of the polarity of each of the tweets. Below, is an example with Compensar tweets:

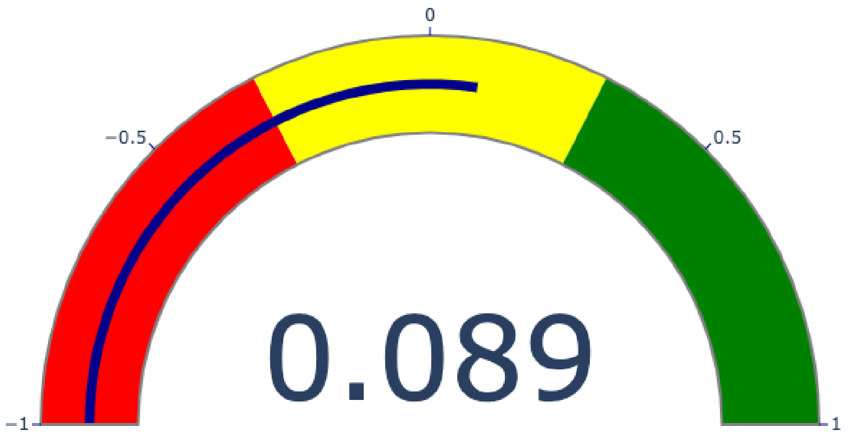


Figure 5.2 Polarity Aggregate Analysis - Gauge Graph.

Deep learning and neural networks:

For each model the dataset was previously splitted in two sets, training and test data with a split fraction of 0.2. During each epoch the data was trained in batches of 32 samples.

<u>Sentiment Analysis</u>

The model classifies the sentiment in two options: Neutral and Negative.

Model 1 Simple network:

Parameters:

Embedding 64

Activation:

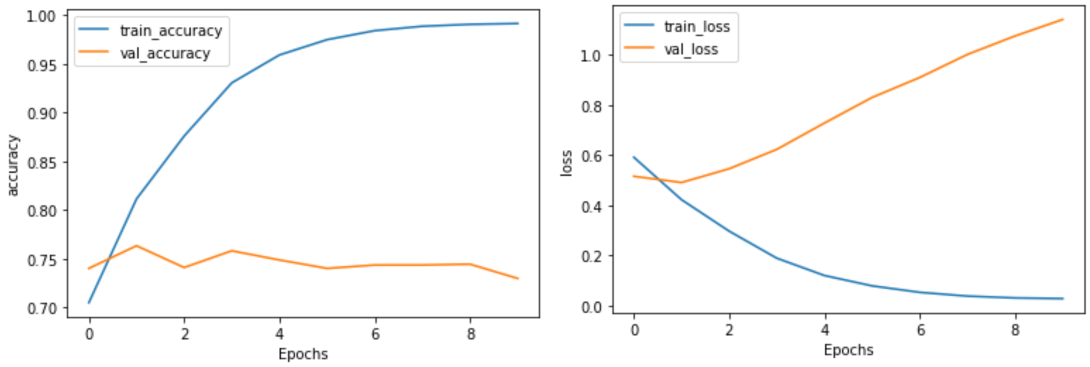
Layers: Relu

Decision layer: Sigmoid

Input length: 60

Epochs: 10

Accuracy: 93% loss: 0.189. Validation 76% and loss: 0.627



In the first model we found overfitting, due to the difference between the accuracy of the train and validation sets, and in the case of the validation, the model didn't have a growing trend.

Model 2 Recursive network LSTM

Parameters:

Embedding 10000

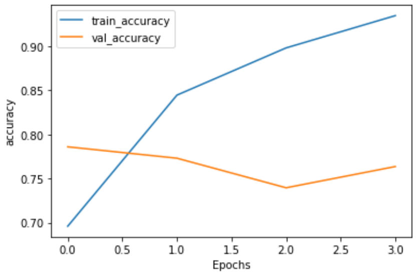
Activation:

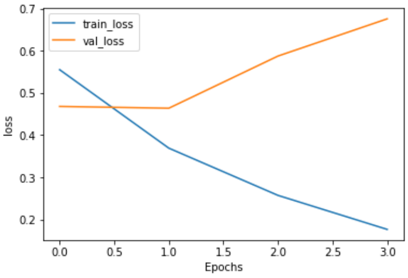
Decision layer: Sigmoid

Input length: 60

Epochs: 10

Accuracy: 93% loss: 0.176. Validation 76% and loss: 0.67





With this model we still have overfitting, and the loss of the validation set increases a lot.

Model 3 Regularization Dropout

Parameters:

Embedding 64

Activation:

Layers: Relu

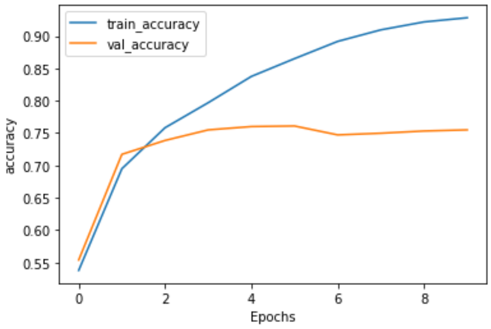
Decision layer: Sigmoid

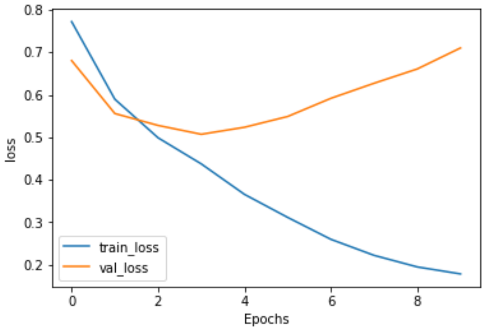
Input length: 60

Epochs: 10

Dropout: 20%

Accuracy: 86.5% loss: 0.31. Validation 76% and loss: 0.549





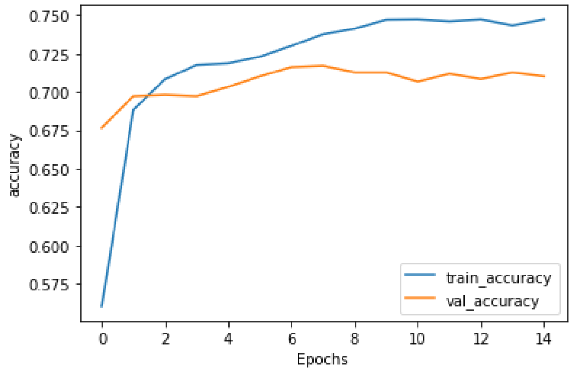
This model presents a better performance, the curves are very close until the epoch 3 and the loss decreases.

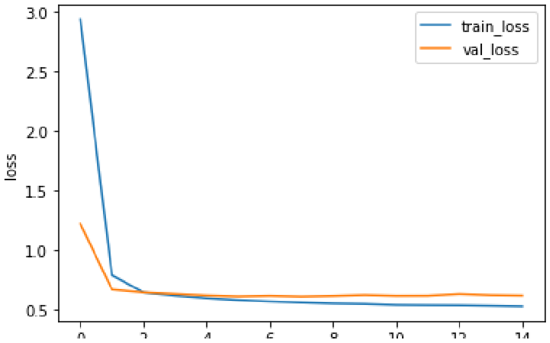
Model 4 Regularization L1

Parameters:

- Embedding 64
- Activation:
  - Layers: Relu
  - Decision layer: Sigmoid
- Input length: 60
- Epochs: 15
- L1: 0.01

Accuracy: 74.3% loss: 0.54. Validation 71% and loss: 0.61

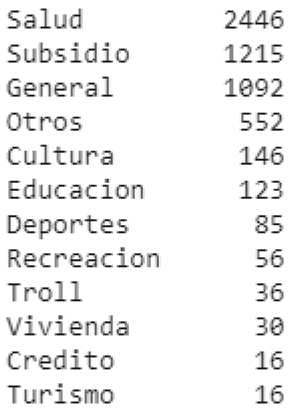




The behaviour of the accuracy and loss in the train and validation sets are closer, being a good model

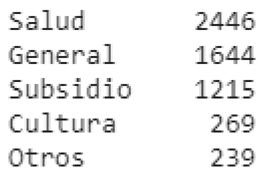
Service Analysis

The Compensation Funds have different services that we manually classified as shown:



```
Salud          2446
Subsidio       1215
General        1092
Otros           552
Cultura         146
Educacion       123
Deportes         85
Recreacion       56
Troll            36
Vivienda         30
Credito          16
Turismo          16
```

We decided to group some of the services in order to have balance between them having this groups:



```
Salud       2446
General     1644
Subsidio    1215
Cultura      269
Otros        239
```

Model 1 Simple network:

Parameters:

Embedding 1000

Activation:

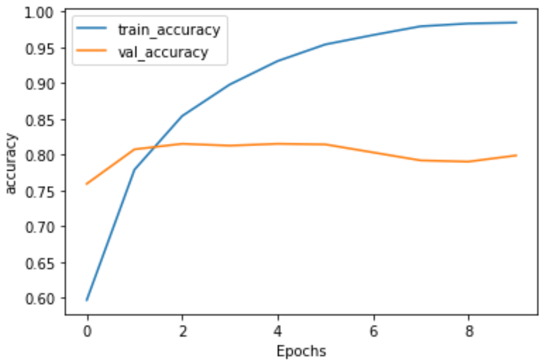
Layers: Relu

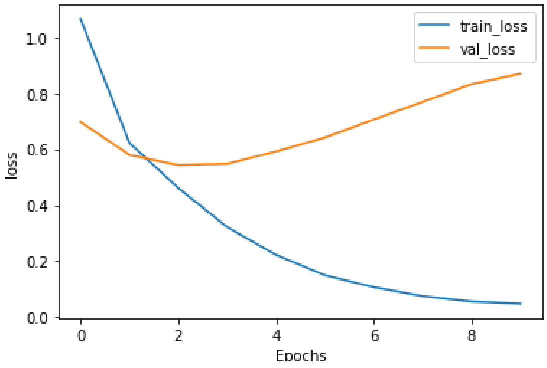
Decision layer: Softmax

Input length: 60

Epochs: 10

Accuracy: 93% loss: 0.22. Validation 81% and loss: 0.59





It's a good model, having an accuracy of 93% and 81%.

Model 2 Recursive network LSTM

Parameters:

Embedding 64

Activation:

Layers: Relu

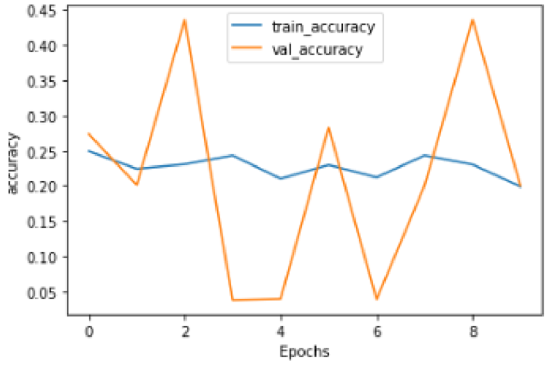
Decision layer: Softmax

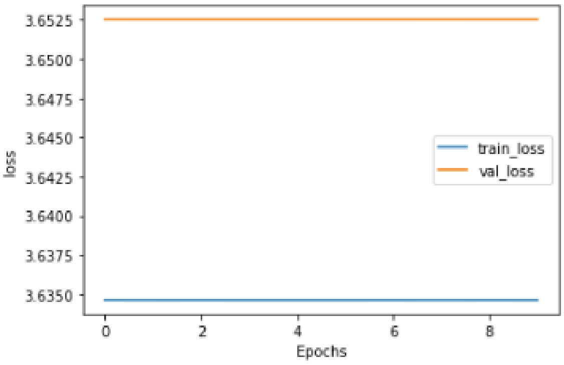
Input length: 60

Epochs: 10

Dropout 20%

Accuracy: 24% loss: 3.64 Validation 20% and loss: 3.65



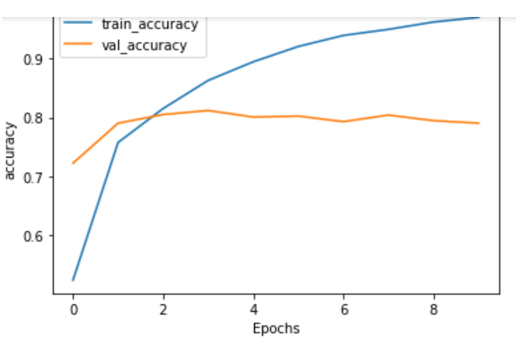


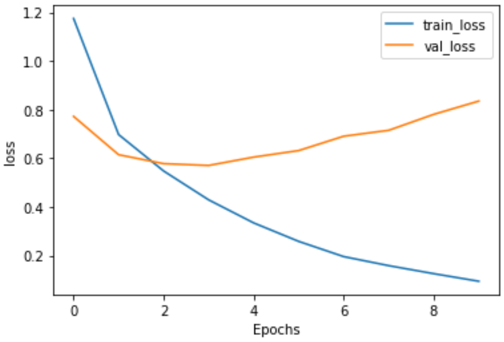
This is a bad model, the accuracy is very low, the losses are constant and very different between them.

Model 3 Regularization Dropout
Parameters:

- Embedding 1000
- Activation:
    - Layers: Relu
    - Decision layer: Sigmoid
- Input length: 60
- Epochs: 10
- Dropout: 10%

Accuracy: 91% loss: 0.28. Validation 81% and loss: 0.60





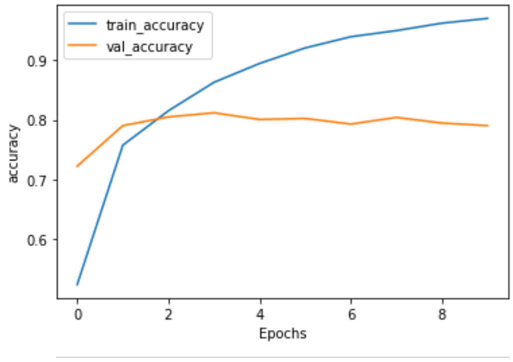
It's a good model, with high accuracy in both train and validation sets.

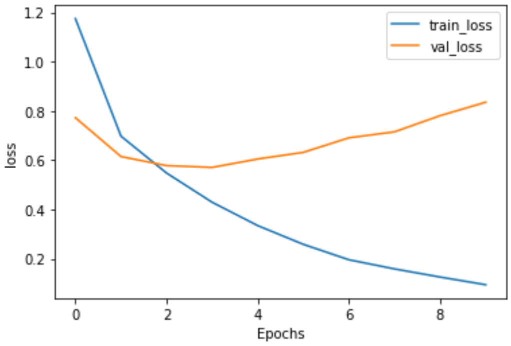
Model 4 Regularization L1
Parameters:

- Embedding 64
- Activation:
    - Layers: Relu
    - Decision layer: Softmax
- Input length: 60

Epochs: 15
L1: 0.01

Accuracy: 87% loss: 0.49. Validation 75% and loss: 0.78





The model is very similar to the Dropout regularization model.

The models selected for the project are:

- Sentiment analysis: Model 2 Recursive network LSTM, with an accuracy of 97%, a loss of 0.176, a validation accuracy of 76% and a validation loss of 0.67.
- Service segmentation: Model 2 Regularization dropout, with an accuracy of 91%, a loss of 0.28, a validation accuracy of 81% and a validation loss of 0.60.

Conclusions:

The model classify each Tweet related to the different compensation funds in positive or negative feeling, and in the different services offered.

We identify that during the Covid, the services more commented are related to health (specifically to the covid test service) and the subsides, specifically the unemployed subside. In general, this two have a negative feeling, while culture has a good feeling across the users, general and others have a positive feeling closer to neutral.

The compensation fund with more affiliates in Colombia is Colsubsidio, followed by Compensar and Comfama, while in Tweets, the found more mentionated is Compensar followed by Comfama and Colsubsidio.

References:

- Theo, he Best Libraries for React I18n, **https://phrase.com/blog/posts/react-i18n-best-libraries/**
- Facebook Open source, React. https://es.reactjs.org/
- Jeanne Ross, ORGANIZATIONAL DESIGN FOR DIGITAL TRANSFORMATION

https://executive-education-online.mit.edu/presentations/lp/mit-organizational-design-for-digital-transformation-online-short-course/?gclsrc=aw.ds&&&ef_id=c:450994018272_d:c_n:g_ti:kwd-303236455599_p:_k:%2Borganizational%20%2Bstructure_m:b_a:107733008849&gclid=Cj0KCQjwgo_5BRDuARIsADDEntQ2DO1DQUesXF8nLMY0SMfeomjJLxH9t5P0wkZ7voVB5V5c3jhnm_gaAuj4EALw_wcB

- Jason Brownlee, Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras. https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/
- Streamlit Inc, Documentation, https://docs.streamlit.io/en/stable/