



Building An Emotionally Annotated Corpus of Investor Tweets

Fernando J. Vieira da Silva *Norton T. Roman*
Ariadne M. B. R. Carvalho

Technical Report - IC-18-05 - Relatório Técnico
March - 2018 - Março

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

The contents of this report are the sole responsibility of the authors.
O conteúdo deste relatório é de única responsabilidade dos autores.

Building An Emotionally Annotated Corpus of Investor Tweets

Fernando J. Vieira da Silva* Norton T. Roman† Ariadne M. B. R. Carvalho‡

Abstract

Emotionally annotated corpora are specially important for training machine learning models for automatic emotion identification, among other applications. However, the task of manually assigning emotions to a corpus carries a high level of subjectivity. In this technical report, we describe the annotation tools and methodology we used for dealing with this challenge when building an emotionally annotated corpus of investor tweets.

1 Introduction

In this report we detail the annotation tools and methodology we followed when building a corpus with manually annotated emotions in tweets from Brazilian stock market investors. In section 2 we define the emotion set we use for annotating tweets, while in section 3 we present the corpus and explain how we collected it, also showing some samples. In section 4 we describe the two annotation tools developed, along with the annotation procedure taken. We present annotators profiles in section 5 and the final conclusion is presented in section 6.

2 Defining Emotion

The concept of emotion is applicable to all evolutionary levels, including animals and humans – in other words, it is considered a psychoevolutionary theory. In this work, we use the Plutchik’s Wheel of Emotions ([4] *apud* [6]), which claims that there are four basic emotional axes, each of them holding two opposite emotions that can vary in intensity and combine to build new emotions. These opposite emotion pairs are: *joy vs sadness*, *anger vs fear*, *trust vs disgust* and *surprise vs anticipation*, forming a wheel, as shown in Figure 1.

Table 1, extracted from [2], shows how this theory can apply to animals at all evolutionary levels. In this table, we can observe the events that lead to each emotion and the actions triggered by them. As for the joy emotion, it is led by the event of a gain of a valued

*Institute of Computing, University of Campinas, Campinas-SP, Brazil

†School of Arts, Sciences and Humanities, University of São Paulo, São Paulo-SP, Brazil

‡Institute of Computing, University of Campinas, Campinas-SP, Brazil

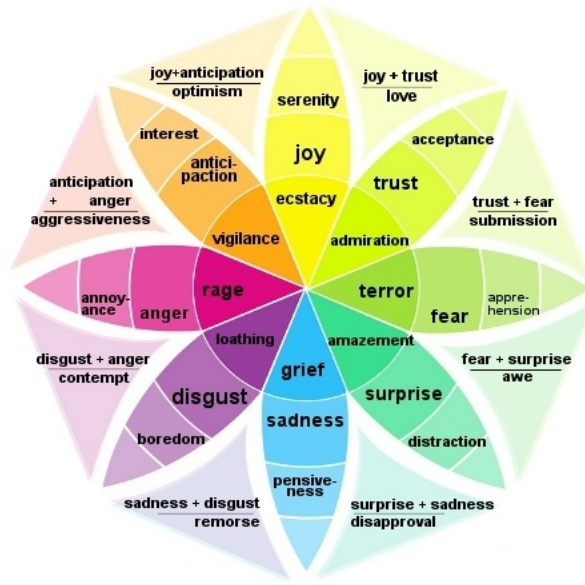


Figure 1: Plutchik Wheel of Emotions – extracted from [5].

object, which comes with the cognitive appraisal of possession, and triggers the behavioral action of retaining or repeating such gain, with the function of gaining resources. When a new member of the group appears, and this event comes along with friendship, then the associated emotion is trust, which leads to partnership, that functions as mutual support for surveillance. The same intuition also applies to other emotions.

Table 1: Survival issues and their relation to the emotions, according to Plutchik’s Wheel of Emotions.

Stimulus Event	Cognitive Appraisal	Subjective Reaction	Behavioral Reaction	Function
gain of valued object	“possess”	JOY	retain or repeat	gain resources
member of one’s group	“friend”	TRUST	groom	mutual support
threat	“danger”	FEAR	escape	safety
unexpected event	“what is it?”	SURPRISE	stop	gain time to orient
loss of valued object	“abandonment”	SADNESS	cry	reattach to lost object
unpalatable object	“poison”	DISGUST	vomit	eject poison
obstacle	“enemy”	ANGER	attack	destroy obstacle
new territory	“examine”	ANTICIPATION	map	knowledge of territory

We chose to work with this theory because it presents the surprise and anticipation emotions, which we believe to be important in the domain of investors' tweets, as stock prices may actually be affected by forecasted or surprising events.

3 Corpus

The corpus is composed of tweets that mention stocks from the IBOVESPA¹ index. These tweets were automatically collected by looking for messages that contain the code of at least one of the 73 stocks that compose the index. Usually, a stock code is a five or six-character alphanumeric string that represents stocks from companies, such as “ABEV3” for AmBev², “PETR3” for Petrobras³, and “BBAS3” for Banco do Brasil.⁴ These codes are so popular among stock market investors that they are commonly used on Twitter as a surrogate for their company names. Some of these tweets can be seen in Figure 2. Tweets were collected during March, 2014, in a total of 7,175 non-repeated tweets.

<i>#bbas3 Depois acho meu post mas ainda aguardo 18,75</i> [<i>#bbas3 I will find my post latter but I still wait for (the stock price to reach) 18.75</i>]
<i>Acordo da ALLL3 melou?</i> [<i>Is the ALLL3 agreement gone?</i>]
<i>Ai meu bolso.... Bbas3 caiu pra c***** hoje</i> [<i>Ouch, my wallet... Bbas3 fell as (expletive) today</i>]
<i>Ativo c/ vol Financeiro Superior a sua MM21-16h:AEDU3 ALUP11</i> ARTR3 BBRK3 BBTG11 BHGR3 BISA3 BRAX11 BRKM5 BRPR3 CCRO3 COCE5 CPFE3 CPLE6 ELET3 [<i>Stocks with financial volume over their MM21-16h:AEDU3 ALUP11</i> ARTR3 BBRK3 BBTG11 BHGR3 BISA3 BRAX11 BRKM5 BRPR3 CCRO3 COCE5 CPFE3 CPLE6 ELET3]
<i>Resumo dos trades:</i> - vendas: csna3 e bbas3 - cancelados: mrve3 e suzb5 - gatilhos ainda armados: itsa4 e brml3 [<i>Trade summary:</i> - sells: csna3 and bbas3 - cancelled: mrve3 and suzb5 - triggers still armed: itsa4 and brml3]
<i>Trade de venda no gráfico semanal acaba de ser acionado em BBAS3</i> [<i>Weekly sell trade sign has just been detected in BBAS3</i>]

Figure 2: Tweets mentioning stocks (company codes in bold).

¹IBOVESPA is the stock market index composed of the main companies' stocks traded at BM&F BOVESPA stock exchange, the most important in Brazil, hosted in São Paulo.

²AmBev is a Brazilian multinational beverage company which produces beers and soft drinks.

³Petrobras is a Brazilian multinational oil state company.

⁴Banco do Brasil is literally translated to “Bank of Brazil”, which is a Brazilian state bank.

This corpus has also some historical importance since this period coincides with several relevant events in recent Brazilian political and economic scenarios, such as the beginning of the *Lava Jato*⁵ operation – a major investigation carried out by the Brazilian Federal Police that investigates many politicians, including former presidents⁶; the scandal of the Pasadena refinery overpaid purchase by Petrobras [1]; the first presidential election polls published and the beginning of the economic recession in Brazil; all of them clearly reflected in the tweets.

<p><i>CASO LAVA JATO - Youssef pode ter atuado em refinaria da Petrobras no Paraná</i> http://t.co/6upiN3J6nR #PETR3 #Petr4 #Lavajato <i>[CAR WASH CASE - Youssef may have acted in Petrobras' refinery in Paraná]</i></p>
<p><i>Eu não sabia que a refinaria americana era uma #BarcaFurada.</i> <i>Sei...</i> <i>#Petrobras sofre e que não acha Graça são os acionistas de #PETR4</i> <i>["I didn't know that American refinery was a bad idea (in a free translation)."]</i> <i>Sure...</i> <i>#Petrobras suffers and #PETR4's shareholders are who does not think it is funny.]</i></p>
<p><i>RT @garimpodeacoes: Conforme esperado ações de estatais sobem qdo Dilma cai nas pesquisas: BBAS3+3,07%, PETR3,+2,30%, PETR4,+2,29%, ELET3,+...</i> <i>[As expected, state companies' stocks raise when Dilma falls in election polls: BBAS3+3,07%, PETR3,+2,30%, PETR4,+2,29%, ELET3,+...]</i></p>

Figure 3: Tweets mentioning recent Brazilian history relevant events (company codes in bold).

4 Annotation Tools

Tweets were manually annotated by volunteers using a tool developed to this specific purpose and following the process defined in [6]. Since annotators may find more than one emotion in a single tweet, every tweet had to be annotated four times, once for each pair of emotions (*joy or sadness, anger or fear, trust or disgust, surprise or anticipation*), so as to prevent tweets from being assigned opposite emotions. Consider, for instance, the following tweet:

Fevereiro 2014: Ainda não é o momento de entrar: VALE5, CSNA3 ou PETR4
Tendência e resultados favoráveis: ESTC3, PSSA3, EMBR3 e WEGE3.
[February 2014: It is not the moment to get in yet (to buy the stocks): VALE5, CSNA3
or PETR4. Trends and good results: ESTC3, PSSA3, EMBR3 and WEGE3.]

During the annotation for “surprise or anticipation”, the following options are available to the annotators:

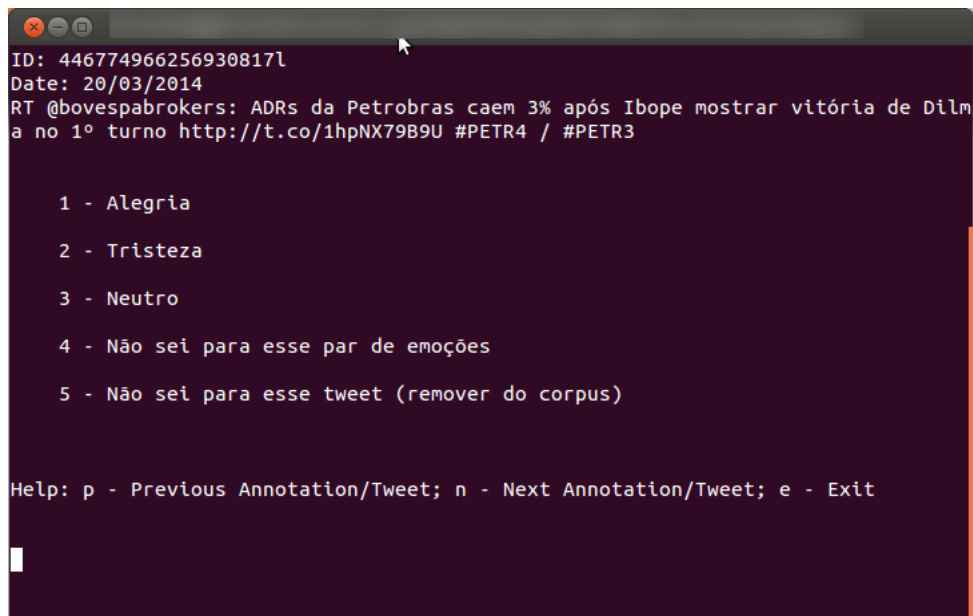
⁵“Car Wash”, in a free translation.

⁶https://en.wikipedia.org/wiki/Operation_Car_Wash.

1. surprise;
2. anticipation;
3. neutral;
4. don't know for this emotion pair;
5. don't know for any pair (leave this tweet out of the corpus).

The first two options will vary for the other three pairs, that is joy or sadness, anger or fear, and trust or disgust; and the last two options will allow the annotators to identify tweets that are difficult to understand or annotate. A tweet is only considered neutral if it is annotated as such for all pairs.

At first, volunteers were asked to annotate 1,000 tweets in exchange for a free online programming course. To do so, they were given a simple command line interface based tool, as shown in Figure 4. This tool strictly followed the annotation process defined by [6]. Therefore, the user has to annotate each pair of emotion at a time using command keys, such as “1” or “2” to indicate one of the two emotions in the pair, the key “3” for neutral, key “4” when the annotator does not know for this pair of emotions and key “5” when he/she does not know how to classify the tweet for any emotion pair and asks to remove the tweet from the corpus. The keys “p” and “n” are used to navigate through previous and next annotation, respectively, and the key “e” is used to exit the tool. In addition, the annotator is able to return later from where he/she stopped. All the tweets are stored locally and periodically uploaded into a server.



```
ID: 4467749662569308171
Date: 20/03/2014
RT @bovespabrokers: ADRs da Petrobras caem 3% após Ibope mostrar vitória de Dilma
a no 1º turno http://t.co/1hpNX79B9U #PETR4 / #PETR3

1 - Alegria
2 - Tristeza
3 - Neutro
4 - Não sei para esse par de emoções
5 - Não sei para esse tweet (remover do corpus)

Help: p - Previous Annotation/Tweet; n - Next Annotation/Tweet; e - Exit
```

Figure 4: Text based annotation tool screen-shot.

The announcement was broadcast directly among friends and groups related to Natural Language Processing/Computational Linguistics and Stock Market, in social networks such as Facebook⁷ and LinkedIn⁸. Only ten volunteers answered it and, from July/2015 to December/2015, only two of them annotated around 1,000 tweets each, whilst the rest dropped out.

We then developed a second version of the annotation tool, in the form of a mobile web based application, So, instead of annotating each emotion in a different screen for the same tweet – as in the first version – this version uses a single screen for the same tweet as seen in Figure 5, showing four rows of four buttons, each row containing one button per emotion in the pair (ie. one pair of emotions per row of buttons) plus one button for “neutral” and another one to indicate the annotator does not know how to classify the tweet according to that pair of emotions. Only when the user chooses at least one option in each of the rows, then he/she can submit the annotation.

Faltam 16 tweets para você ganhar mais um bilhete!

 Ajuda
Tweet ID:
456398217542766592

**\$CSAN3 - Cosan (csan-nm) -
Fato Relevante - Aprovação-
proposta De Incorporação-all
E Rumo**
<http://t.co/ctpNxKDEsC>

 Ajuda

Alegria	Tristeza	NEUTRO	Não sei
---------	----------	--------	---------

 Ajuda

Confiança	Desgosto	NEUTRO	Não sei
-----------	----------	--------	---------

 Ajuda

Raiva	Medo	NEUTRO	Não sei
-------	------	--------	---------

 Ajuda

Antecipação	Surpresa	NEUTRO	Não sei
-------------	----------	--------	---------

 Ajuda

Figure 5: Mobile Based Annotation tool. In the top, there is an indication of the number of tweets still missing to another lottery ticket. Next, there is the tweet id. The tweet’s text is found below in bold. The buttons in green were selected by the user.

This time, however, we asked volunteers to annotate the number of tweets they would

⁷ www.facebook.com

⁸ www.linkedin.com

like and, at every 20 tweets, annotators would receive a ticket, giving them a chance to win a prize in cash of 100.00 Brazilian Real. As tweets are annotated individually, even if the user logs in to annotate a single tweet, it would still be added to the user count and, as soon as the count reaches 20, the annotator gets a new ticket. Finally, and as another incentive to keep annotators in the experiment, at every annotated tweet, they were informed about the amount of tweets remaining for another lottery ticket, as shown in Figure 6.

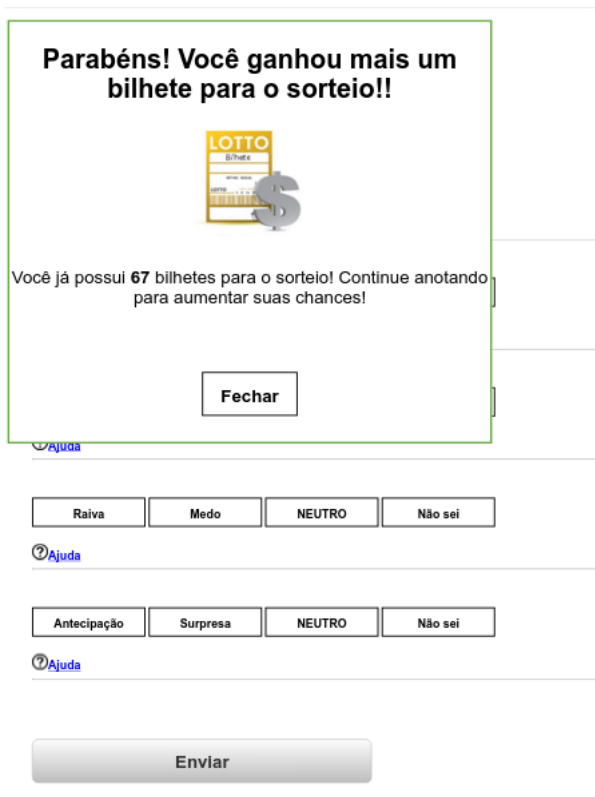


Figure 6: The message indicating the gain of a new lottery ticket. In bold: “Congratulations! You earned a new lottery ticket!” in a free translation. Continuing below the figure: “You already have ’n’ lottery tickets! Keep annotating to increase your chances of winning!”.

We announced this new experiment in the same channels as we did before, also doing it in the students mailing list of the University of Campinas, one of the largest universities in Brazil. By the end of the annotation period, from December 27th, 2016 to February 6th, 2017, three random tickets were drawn and the prizes paid. As a result, we found 442 volunteers and we obtained 4,517 tweets analyzed by at least three people.

The application developed for the annotation process was designed so as to try to maximize the number of tweets annotated by at least three different people. To do so, and since it was a web-based tool, at the server side the algorithm for choosing a tweet to be annotated was as follows: 1) Pick a random tweet among the ones already annotated by two people; 2) If none is found, then select a random tweet in the subset of tweets with a

single annotation; 3) In case there is still no tweet available, then pick a random tweet with no annotation; 4) As a last option, if all the tweets already have three annotations, then pick one at random. In doing so, at every moment, the system has the maximum number of annotations for each tweet.

Finally, since when a tweet is assigned to a user it is marked as reserved, the number of reserved tweets is also considered when choosing a tweet for annotation, which means a tweet annotated by one person and reserved for another person (i.e. he/she is currently annotating it) falls into the option 1 above. However, if that person does not submit within 15 minutes, the server assumes him/her to have dropped out and revokes the tweet reservation, so that the tweet may be passed on to some other annotator – if this user still submits his/her annotation after this period, however, it will be registered anyway. In the worst case we will end up with a tweet with four annotations (or more) whilst there are still tweets with less than three annotations in the database. An additional possible advantage of this approach is the fact that, since tweets are randomly picked, each user follows a different sequence of tweets when annotating them, thereby reducing any bias introduced by some specific sequence.

5 Participants

As it can be seen in Figure 7, annotators were evenly distributed between women and men and their ages vary between 18 and 60 years old, being 52.5% in the range starting from 18 to 25 years old; 27.6% younger than 30 years old; 16.7% between 31 and 40 years old, 2.7% from 41 to 50 years old and only 0.5% older than that.

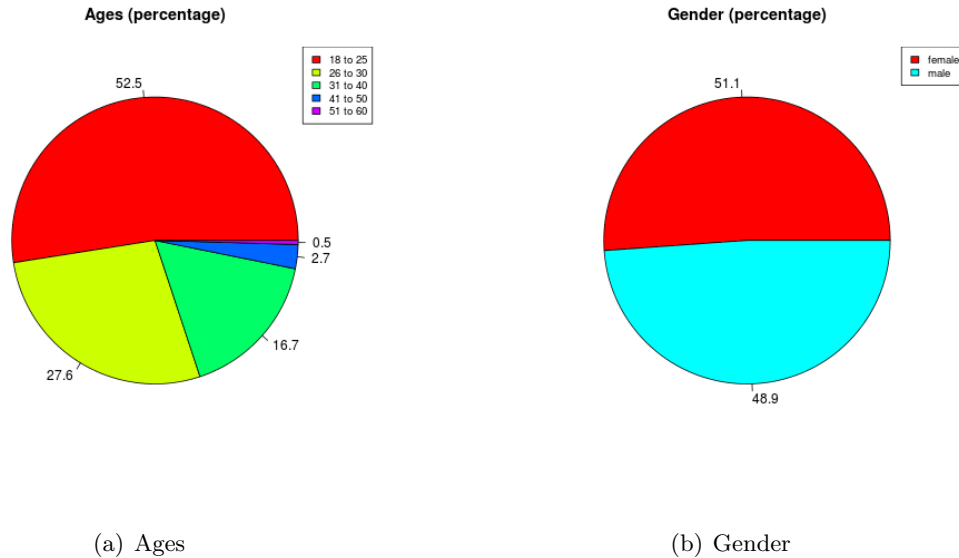


Figure 7: Annotators personal profiles.

When we look at their academic background, we can observe that most of them (44.6%)

were undergraduate students, as illustrated in Figure 8. However, more than half of the annotators were graduated, more specifically, 22.2% were graduates, 21.3% held a master degree and 10.9% a doctorate degree. This considerably high education level may be explained by the fact that the majority of volunteers attended the announcement made in the university mailing list. As for their subject areas of study (either they have been studying or have already graduated), the largest part of them is composed by math and engineering (47.7%), followed by humanities (23.5%) and biological sciences (22.4%). Another 5.2% is from social sciences and 1.1% don't have an area – the same amount of people who graduated in high school only.

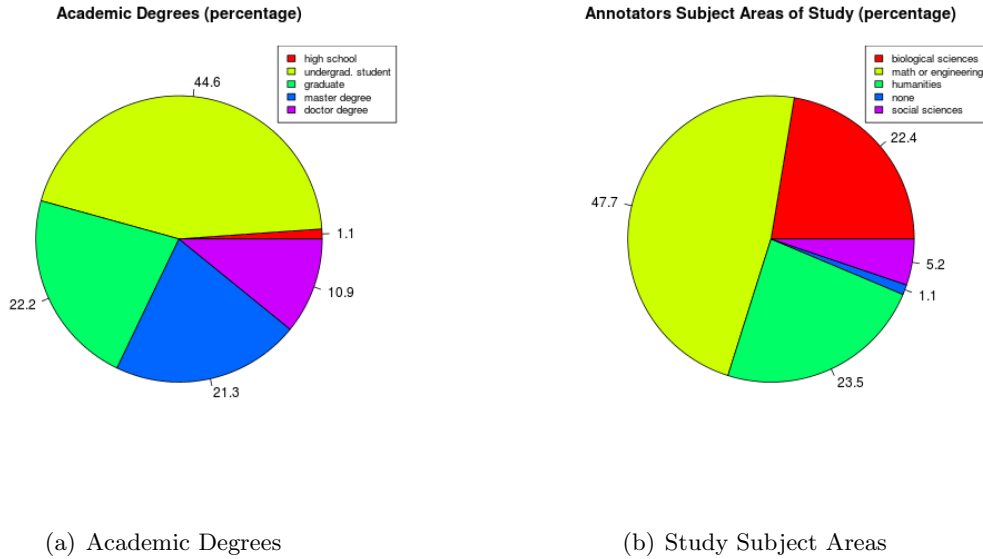


Figure 8: Annotators academic profiles.

We also inquired their experience with stock market, as it might have helped them to understand tweet contents whenever they contain context specific terms and slangs. However, only 14.5% had previous experience trading in the stock market and the same amount has previous knowledge in Technical Analysis [3]. These numbers are shown in Figure 9.

6 Conclusion

In this work, we presented the annotation tools and methodology used for building an emotionally annotated corpus of tweets from Brazilian stock market investors. The corpus was built by downloading tweets containing stock codes and the emotion set used for annotating was also defined.

We also described in details the two annotation tools developed for the task and we showed the participants' profiles.

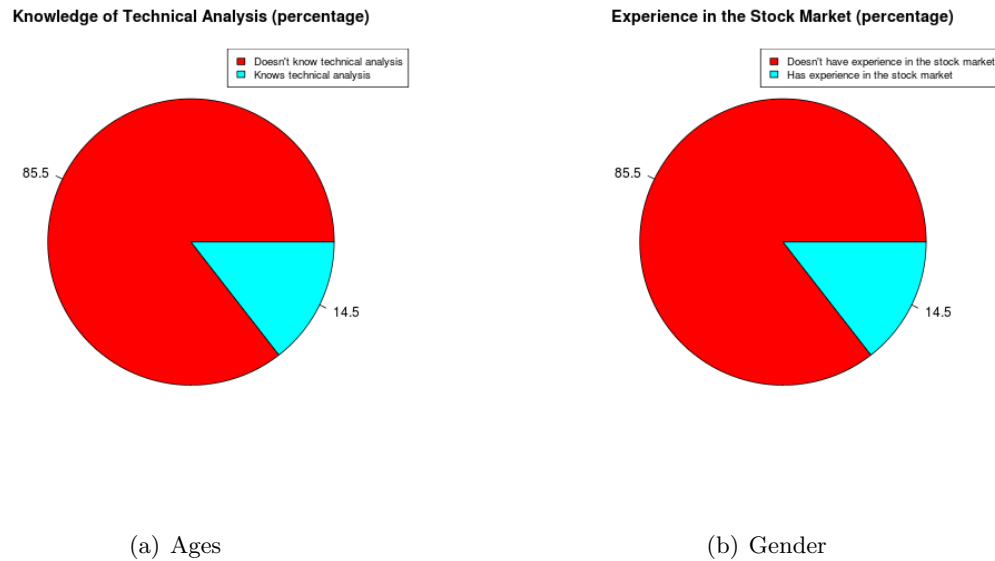


Figure 9: Experience with stock market.

References

- [1] Mihir Zaveri Susan Carroll and Ben Tavener. Brazilian scandal centers on purchase of Pasadena refinery. <http://www.houstonchronicle.com/news/houston-texas/houston/article/Brazilian-scandal-centers-on-purchase-of-Pasadena-6618757.php>, 11 2015. Accessed at 05-29-2017.
- [2] Markus Drew and Mathias Krohn. Robert Plutchik's Psychoevolutionary theory of basic emotions. <http://www.adliterate.com/archives/Plutchik.emotion.theorie.POSTER.pdf>, February 2007. Accessed at: 03-28-2017.
- [3] Andrew W. Lo, Harry Mamaysky, and Jiang Wang. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *Journal of Finance*, 40:1705–1765, 2000.
- [4] Robert Plutchik and Henry Kellerman. *Emotion: theory, research and experience*. Acad. Press, 1986.
- [5] David Pollack. Plutchik's Wheel of Emotions Cheat Sheet. <https://www.cheatography.com/davidpol/cheat-sheets/plutchik-s-wheel-of-emotions/>. Accessed at 03-28-2017.
- [6] Jared Suttles and Nancy Ide. Distant supervision for emotion classification with discrete binary values. In *Computational Linguistics and Intelligent Text Processing*, pages 121–136. Springer, 2013.