Statistica & Applicazioni Vol. XX, n. 1, 2022, pp. 61-78
Published online June 2023

EXPLORING FINANCIAL MICROBLOGS: ANALYSIS OF USERS' TRADING PROFILES WITH MULTIVARIATE STATISTICAL METHODS

Matteo Ventura*
Riccardo Ricciardi*

SUMMARY

StockTwits is a Social Media focused on finance that is receiving increasing attention from finance experts and enthusiasts. In this work, StockTwits' users are studied considering some of their self-declared characteristics, such as trading experience, holding period of the stocks, and trading approach. A Correspondence Analysis is carried out to investigate the relationships among these characteristics, the Simple Correspondence Analysis is applied to study the relationships between the approach and the holding period. The association between these variables and the experience is studied with the Multiple Correspondence Analysis. In the end, a cluster analysis carried out with hierarchical clustering is used to study the structure of the StockTwits community on the basis of the same characteristics. The analyses highlighted that the way users label their own approach and primary holding period reflect the objective relation linking technical strategy with short-term investments and fundamental approach with long-term ones. Moreover, it showed a weak relation of the experience in trading with the other features, configuring self-reported experience as a more cross-sectional characteristic.

Keywords: Cluster Analysis, Correspondence Analysis, Social Media, StockTwits.
DOI: 1026350/999999_000051
ISSN:1824-6672 (print) 2283-6659 (digital)

1. INTRODUCTION

The expansion of the Internet and Social Media platforms (e.g., microblogs) has enabled the availability of a lot of data at a low cost which attracted the attention of the scientific community, which is interested in using these data for a lot of purposes.

* Dipartimento di Economia e Management - Università di Brescia - C.da Santa Chiara, 50 - 25122 BRESCIA (e-mail: ✉ matteo.ventura@unibs.it, r.ricciardi@unibs.it).

Authors' contribution: Conceptualization: Riccardo Ricciardi, Matteo Ventura; Methodology: Matteo Ventura; Data pre-processing: Riccardo Ricciardi; Formal analysis and investigation: Matteo Ventura; Writing - original draft preparation: Riccardo Ricciardi, Matteo Ventura; Writing - review and editing: Riccardo Ricciardi, Matteo Ventura.

The universe of Social Media platforms can be divided into different communities. In this paper, the focus is specifically on the financial community, which has grown in the last years and thus is potentially representative of all investors (Sabherwal, Sarkar and Zhang, 2011).

One of the biggest Social Media platforms used by the financial community is StockTwits (stocktwits.com), which kindly provided the data analyzed in this work. StockTwits is a specialized financial microblog founded in 2009, to let investors, traders, and finance enthusiasts share their ideas about the stock markets. It is a microblogging platform, where people's activity consists of posting tweets, i.e. messages, images, videos, and similar contents, and interacting with other users' posts.

Investors can be classified according to different personal characteristics. Specifically, on StockTwits, users can provide information about their experience, the holding period of their stocks, and their approach in trading.

To the best of our knowledge, the literature about the financial community on social media mainly focuses on users' sentiment about stock markets and its relation with different market variables, using both Twitter (Oliveira, Cortez and Areal, 2017; Umar, Gubareva, Yousaf and Ali, 2021) and StockTwits data (Oliveira, Cortez and Areal, 2013; Audrino, Sigrist and Ballinari, 2020). This work, instead, aims at analyzing the profiles of the users and searching for relationships and patterns among these trading features.

To carry out this analysis, two statistical techniques were applied: Correspondence Analysis and Cluster Analysis.

The rest of the paper is organized as follows. Section 2 presents the data. Next, Section 3 presents the statistical methods applied to analyze the data. Finally, Section 4 discusses the main results and in Section 5 the main conclusions are drawn.

2. DATA

On the StockTwits platform, users may fill in their profile by writing a *bio*, i.e. a short self-description, with a maximum length of 250 characters; and by indicating their trading characteristics, i.e. their experience as traders, their approach in trading, and their primary holding period.

This study uses the same data as in Ricciardi (2022). All the tweets from 2010 to 2021 have been accessed, keeping all the available information about the posting users; then the cleaning and pre-processing phase included the following steps:

1. keeping only tweets with both bio and trading information;
2. retaining the first record per user, to store only the first completion of the user profile;
3. using the *Google Compact Language Detector Algorithm*[1] to retain users with an English-written bio;
4. keeping users whose bio had at least 6 words[2].

[1] https://cloud.r-project.org/web/packages/cld3/index.html.

[2] The minimum length was detected as the 25-th percentile of the number of words per bio.

Eventually, this process yielded a dataset of 23453 StockTwits users with a textual self-description and self-declared trading characteristics.

2.1 *Trading profiles*

As mentioned above, three self-declared trading characteristics were considered: *experience*, *approach*, and *holding period*.

StockTwits asks the user to indicate its *experience* based on an ordinal scale: *Novice*, *Intermediate*, *Professional*.

By indicating their primary *holding period*, users provide information about the amount of time they usually hold an investment, i.e. the period between the purchase and the sale of a security. The variable *Holding Period* has four categories that can be sorted according to an increasing holding period: *Day Trader*, *Swing Trader*, *Position Trader* and *Long Term Investor*. Day traders are the investors that try to exploit market inefficiencies and buy and sell securities within the same day. On the opposite side, long-term investors hold the bonds for more than one year. This category, therefore, tolerates the risk and waits for the premium. Then, there are *position traders* and *swing traders* that have a *holding period* ranging from weeks to months, and from some days to some weeks, respectively.

With regard to *Approach*, the platform lets users choose among *Fundamental*, *Growth*, *Value*, *Technical*, *Momentum*, and *Global Macro*. Observing the relative frequencies of the categories, it was noticed that the category *Global Macro* had a low frequency (3.08%) compared to the other categories. In Correspondence Analysis, small categories have too much influence on the results (Husson, Lê and Pagès, 2011), therefore this class was removed from the dataset.

Despite users can flag only one option, these categories do not detect only alternative strategies. Indeed, on the one hand, with the purpose of identifying future trends, *technical analysis* studies variations in price, volume, and open interest, by mainly using charts (Murphy, 1999). In this context, some technical traders focus on *momentum* indicators measuring the velocity of price changes, namely how ‘persistent’ a certain price is, and then decide whether to trade a stock or not. On the other hand, *fundamental analysis* attempts to identify overvalued and undervalued securities through the analysis of industry and companies’ income statements and characteristics. In particular, *value* investors use the aforementioned instruments of analysis to compute the intrinsic value of a stock, in order to compare it with its market price, and then buy and hold it until the former value is greater than the latter (Graham and Daniels, 2015). Differently, by adopting a *growth* strategy, other traders invest in companies that are in the *growth* phase of their life cycle, i.e. before their *maturity*, and whose growth rate is greater than the average of their industry (Price, 1973).

Due to the timing release of the means used for their analysis, fundamental and technical investors tend to invest in middle/long-term and short-term, respectively (Price, 1973; Murphy, 1999; Graham and Daniels, 2015). Therefore, it would be interesting to investigate whether StockTwtits users’ profiles reflect this correspondence.

In Table 1 a summary of the absolute and relative frequencies of the categories of each variable is reported.

TABLE 1. - *Absolute and relative frequencies of the categories for each variable*

Variables	**Abs. Freq.**	**% Freq.**
Approach		
Fundamental	2854	12.56
Growth	3791	16.68
Value	2127	9.36
Momentum	4018	17.68
Technical	9940	43.73
Holding period		
Day Trader	5185	22.81
Swing Trader	9360	41.18
Position Trader	3978	17.50
Long Term Investor	4207	18.51
Experience		
Novice	5060	22.26
Intermediate	10209	44.91
Professional	7461	32.82

3. STATISTICAL METHODS

In this section, the statistical techniques applied for analyzing the data are presented. In order to study the associations between the variables *Experience*, *Holding period*, and *Approach*, Correspondence Analysis was applied. Indeed, Cluster Analysis was used to study the clustering structure of StockTwits user profiles according to the three considered variables.

3.1 *Correspondence analysis*

Correspondence Analysis (CA) (Benzécri, 1973) is a multivariate statistical technique that aims at studying the relationships between the categories of two or more categorical variables (Cazes, 2014; Kamalja and Khangar, 2017).

CA can be Simple or Multiple; Simple Correspondence Analysis (SCA) is applied to study the relationships between two categorical variables, Multiple Correspondence Analysis (MCA), instead, is applied when the object of interest is more than two categorical variables (Bolasco, 2002).

CA relies on the general framework of the Exploratory Factor Analysis (EFA), which is a set of techniques, introduced at the beginning of the XXth century by Pearson (1901) and Spearman (1904), which are usually applied to investigate the underlying structure of a large set of variables: it reduces the set into a smaller one (Taherdoost, Sahibuddin and Jalaliyoon, 2022).

The aim of these methods is to reduce the multidimensionality of the data projecting the cloud of points on a subspace by minimizing the implicit distortion of the projection. These subspaces are generated by new variables, called dimensions, components, or factors, that are uncorrelated among them (Bolasco, 2002).

The general framework on which both SCA and MCA are based requires the triple of matrices $(\mathbf{X}, \mathbf{W}, \mathbf{M})$ which respectively are the data matrix to be analyzed, the diagonal matrix of the weights w_i, and the symmetric positive-definite matrix of the metric.

The matrix $\mathbf{M}$ contains information about the dissimilarity of the categories in the space. CA requires a specific metric: the chi-square (χ^2) distance, which measures the dissimilarity of the categories considering both the distance of the categories and the importance of each category.

The subspaces are found through the following maximization problem which can be solved through eigen-equations:

$$\begin{cases} \max\limits_{u} \mathbf{u}'\mathbf{M}\mathbf{X}'\mathbf{W}\mathbf{X}\mathbf{M}\mathbf{u}, \\ \mathbf{u}'\mathbf{M}\mathbf{u} = 1 \end{cases} \tag{1}$$

The problem results in a set of eigenvectors $\mathbf{u}$ which represent the axes of the subspace, and eigenvalues λ which represent the inertia, i.e. the information, retained by the subspace, while the sum of eigenvalues represents the total inertia of the cloud of data points.

3.1.1 *Simple correspondence analysis*

Simple Correspondence Analysis is based on a contingency table $\mathbf{N}$, which can be analyzed from two points of view: considering row or column profiles. In both cases, the categories of one of the variables are described through the categories of the other variable (Bolasco, 2002).

The joint absolute frequencies of the contingency table $\mathbf{N}$ are denoted by n_{ij}, the column and row absolute marginal frequencies are respectively denoted by $n_{.j}$ and $n_{i.}$, and the grand total of observations is denoted by n.

SCA allows the simultaneous study of the cloud of the r row profiles in $\mathbb{R}^c$ and the cloud of the c column profiles in $\mathbb{R}^r$. SCA searches for the optimal representation of r and c points in low-dimensional Euclidean space (Cazes, 2014).

The row profiles $\mathbf{P}_r$ are defined as:

$$\mathbf{P}_r = \mathbf{W}_r^{-1}\mathbf{N},$$

where $\mathbf{W}_r = diag(\ldots n_{i.} \ldots)$ and the generic element of $\mathbf{P}_r$ is $(n_{ij} \,/\, n_{i.})$.

Similarly, the column profiles $\mathbf{P}_c$ are defined as:

$$\mathbf{P}_c = \mathbf{W}_c^{-1}\mathbf{N}',$$

where $\mathbf{W}_c = diag(\ldots n_{.j} \ldots)$ and the generic element of $\mathbf{P}_c$ is the relative frequency $(n_{ij} \,/\, n_{.j})$.

Therefore, the triple of matrices $(\mathbf{X}, \mathbf{W}, \mathbf{M})$ is modified as follows:

- for the analysis of the row profiles in $\mathbb{R}^c$:
  $\mathbf{X} = \mathbf{P}_r(r \times c) = \mathbf{W}_r^{-1}\mathbf{N}$;

$\mathbf{M} = \mathbf{M}_r = n\mathbf{W}_c^{-1}$, is a $c \times c$ matrix that contains the inverse of column marginal frequencies;
$\mathbf{W} = n^{-1}\mathbf{W}_r$ is a diagonal $r \times r$ matrix containing the row marginal frequencies.

- For the analysis of the row profiles in $\mathbb{R}^r$:
  $\mathbf{X} = \mathbf{P}_c\ (c \times r) = \mathbf{W}_c^{-1}\mathbf{N}'$;
  $\mathbf{M} = \mathbf{M}_c = n\mathbf{W}_r^{-1}$, is a $r \times r$ matrix that contains the inverse of the row marginal frequencies;
  $\mathbf{W} = n^{-1}\mathbf{W}_c$ is a diagonal $c \times c$ matrix containing the column marginal frequencies.

The optimization problem in (1) can be adapted to SCA by replacing the triple of matrices with the new matrices shown above.

There exist perfect symmetry between row-profile analysis and column-profile analysis and the points of the two variables can be represented on the same plane (Husson *et al.*, 2011).

3.1.2 *Multiple correspondence analysis*

MCA can be considererd as an extension of SCA to the case of more than two categorical variables (Bolasco, 2002; Husson *et al.*, 2011; Husson and Josse, 2014).

The analysis can be carried out starting from the $n \times p$ indicator matrix of dummy variables $\mathbf{Z}$, where n is the number of observations and p is the sum of the categories of all the variables (Bolasco, 2002).

The s categorical variables can be crossed in a contingency hypercube (Bolasco, 2002) that can be represented through the Burt table $\mathbf{B} = \mathbf{Z}'\mathbf{Z}$, that contains all the faces of the contingency hypercube. Namely, the Burt table represents all the pairwise associations between the variables, including on the diagonal the associations between each variable and itself. In this table, only the information about the categories is available, while the information about the individuals is not present (Husson and Josse, 2014). MCA can be also considered as the SCA of the Burt Table (Cazes, 2014).

Given the $p \times p$ matrix $\mathbf{D}$ which is the diagonal matrix containing the same diagonal elements of the Burt matrix, like in SCA, the following triples of matrices can be defined according to the space in which the analysis is carried out. The analysis can be performed considering either the observations, represented in the space $\mathbb{R}^p$, or the categories, represented in the space $\mathbb{R}^n$.

For the row profiles whose cloud of points belongs to the space $\mathbb{R}^p$:

- $\mathbf{X} = \mathbf{P}_r = s^{-1}\mathbf{Z}$ is the $n \times p$ matrix of row profiles;
- $\mathbf{W} = \mathbf{W}_r = n^{-1}\mathbf{I}$ is the $n \times n$ weights diagonal matrix;
- $\mathbf{M} = \mathbf{M}_r = ns\mathbf{D}^{-1}$ is the $p \times p$ matrix of the metric that is obtained through the χ^2 distance.

For the column profiles whose cloud of points belongs to the space $\mathbb{R}^n$:

- $\mathbf{X} = \mathbf{P}_c = \mathbf{D}^{-1}\mathbf{Z}'$ is the $p \times n$ matrix of column profiles;

- $\mathbf{W} = \mathbf{W}_c = (ns)^{-1}\mathbf{D}$ is the $p \times p$ weights diagonal matrix;
- $\mathbf{M} = \mathbf{M}_c = n\mathbf{I}$ is the $n \times n$ matrix of the metric that is obtained through the χ^2 distance.

The maximization problem to solve in order to find the new dimensions is the problem (1) and the matrix to be diagonalized can be found by replacing the matrices introduced above in the original problem.

3.1.3 *Factor retention*

Once the dimensions have been obtained, the subsequent step is to decide how many components to retain in order to represent the cloud of points through a subspace considerably smaller than the original one but that also ensures a good representation (Taherdoost *et al.*, 2022).

The scree plot is a useful tool to choose the number of dimensions to retain. It is a visual method and it is a widely applied approach. The number of factors to retain is decided by examining a plot that represents the values of the eigenvalues of the dimensions and looking for a point at which the proportion of inertia explained by each subsequent dimension drops off. This is often referred to as an elbow in the scree plot (Jolliffe, 2002).

3.1.4 *Interpretation of the results*

Since the representation of a cloud of data points through factors and a subspace is an approximation, it is important to check the optimality of the results.

Three indexes can be used to assess the goodness of the results:

1. the proportion of inertia explained by a factor α, τ_α, defined as the ratio between the eigenvalue of the dimension and the sum of the eigenvalues:

$$\tau_\alpha = \frac{\lambda_\alpha}{\sum_{\alpha=1}^{p} \lambda_\alpha}. \tag{2}$$

2. CTR index, that is the contribution of any point i to the inertia reproduced on any axis α (Bolasco, 2002):

$$CTR_\alpha(i) = \frac{w_i c_\alpha^2(i)}{\lambda_\alpha}, \tag{3}$$

   where $c_\alpha^2(i)$ denotes the coordinate of the projection of the ith data point on the factor α.
   The sum of the index for all the dimensions is 1.

3. The QLT index assesses how well element i is represented on axis α. It calculates the factor's share in reconstructing the element and thereby determines the factor's

effectiveness in representing it (Bolasco, 2002). The index, for the individual x_i on the dimension $\Delta\alpha$, is expressed as follows:

$$QLT_\alpha(i) = \cos^2\theta_{x_i,\Delta\alpha} = \frac{\left\|x_{i,\Delta\alpha}\right\|_{\mathbf{M}}}{\left\|x_i\right\|_{\mathbf{M}}} = \frac{c_\alpha^2(i)}{\sum_{\alpha=1}^{p} c_\alpha^2(i)}. \tag{4}$$

If QLT over an axis is near 1, the point is well represented; therefore, if the ratio is equal to 1, it means that the point x_i is on the axis (Cazes, 2014).

3.2 *Cluster analysis*

Cluster Analysis is a multivariate statistical technique that aims at grouping observations such that those in the same cluster are more similar than observations in other groups (James, Witten, Hastie and Tibshirani, 2013; Kauffman and Rousseeuw, 2009). This objective is reached through several types of algorithms that differ in the way they create a cluster.

There are two main families of clustering algorithms: partitioning and hierarchical algorithms. Partitioning methods classify the data into a predetermined number of k clusters; while hierarchical methods do not require to specify the number of clusters *a priori*. In this work, Agglomerative Nesting (AGNES) (Kauffman and Rousseeuw, 2009) is applied, which is an algorithm belonging to the family of hierarchical methods. The algorithm starts with all the objects apart, i.e., with n clusters containing one observation each. At each step two objects are merged until only one is left (James *et al.*, 2013).

3.2.1 *Inter-cluster and intra-cluster dissimilarity*

To group observations that are most similar, some algorithms, like AGNES, rely on the concept of dissimilarity, which is linked, although antithetical, to the concept of similarity (Kauffman and Rousseeuw, 2009).

In order to determine the dissimilarity between observations which are characterized by both numerical and categorical variables, the Gower Coefficient (Gower, 1971) is used. Dissimilarity between observations i and i' belonging to a data set with p mixed variables is defined as:

$$d_{GW}(i,i') = \frac{\sum_{j=1}^{p} \delta_{ii'}^{j} d^j(i,i')}{\sum_{j=1}^{p} \delta_{ii'}^{j}}, \tag{5}$$

where the indicator $\delta_{ii'}^{j}$ is equal to 1 when both the measurements x_{ij} and $x_{i'j}$ for the variable j are non-missing, otherwise it is equal to 0, therefore a pairwise deletion is applied. The number $d^j(i,i')$ is the contribution of the jth variable to the whole dissimilarity between objects i and i'.

There are different methods to compute the element $d^j(i,i')$ according to the type of variables. For nominal and binary variables it is defined as:

$$d^j(i,i') = \begin{cases} 1 \text{ if } x_{ij} \neq x_{i'j} \\ 0 \text{ if } x_{ij} = x_{i'j} \end{cases} \tag{6}$$

Ordinal and interval-scaled variables, instead, are first replaced by their rank and then the following formula is applied:

$$d^j(i,i') = \frac{\left|x_{ij} - x_{i'j}\right|}{R_j} \tag{7}$$

where R_j is the range of variable j defined as $\{\max_{(h)} x_{hj} - \min_{(h)} x_{hj}\}$, where h runs over all non-missing objects for variable j (Rousseeuw, 1987).

Hierarchical algorithms also require information about the inter-cluster dissimilarity, i.e., the dissimilarity between clusters. The three more common types of linkage that fits for AGNES algorithm are: *complete*, *group average* and *single*. Complete linkage defines the dissimilarity between the observations belonging to two clusters as the largest dissimilarity between the observations belonging to the two sets. Group average linkage defines the dissimilarity between the observations belonging to two sets as the average of all pairwise dissimilarities of observations belonging to the clusters. Finally, single linkage defines the dissimilarity between the observations belonging to two clusters using the definition in pure mathematics, where the distance between two sets is the infimum of all pairwise distances (James *et al.*, 2013; Rousseeuw, 1987).

3.2.2 *Cluster validation and stability*

Cluster analysis is often carried out in an exploratory manner and the found patterns are not necessarily meaningful. Therefore, it is important to validate the results, i.e., evaluate their goodness.

Aschenbruck and Szepannek (2020) tried to extend the definition of some of the most common indices originally defined for Cluster Analysis of numerical data to mixed-type Cluster Analysis. In the following, some of them are exposed.

Cindex (Everitt, Landau, Leese and Stahl, 2011), *McClain index* (McClain and Rao, 1975) and *Ptbiserial index* (Linacre and Rasch, 2008). These indices can be generalized and extended to mixed-type clustering because they are based on the distances between the objects to be clustered, therefore the distance matrix is computed through the Gower distance (5) can be used to compute them.

The *silhouette index* (Rousseeuw, 1987) gives a measure of how similar an object is to its own cluster compared to other clusters. The index can be computed for all the observations. Given the silhouette of all the observations, the average silhouette width of each cluster and the average silhouette width of the entire data set can be computed.

TABLE 2. - *Table of validation indices*

Index	**Index range**	**Optimality criterion**
Cindex	$[0,1]$	minimum
McClain index	$[0,+\infty)$	minimum
PtBiserial index	$(-\infty,+\infty)$	maximum
Silhouette index	$[-1,+1]$	maximum
Dunn index	$[0,+\infty)$	maximum

The *Dunn index* (Dunn, 1974) is defined as the ratio between the distance between the closest points belonging to two clusters and the maximum diameter among all the clusters.

In Table 2 the validation indices introduced above are shown with their range and their optimality criterion.

Another aspect to take into consideration when evaluating the results is the stability of the clusters, which is the tendency of a cluster not to disappear if the data set changes in a non-essential way. A bootstrap procedure, proposed by Hennig (2007), is applied to assess the stability of the clusters: the Jaccard coefficient (Jaccard, 1912) is bootstrapped and used to measure the similarity between clusters at each iteration.

4. RESULTS

In this section, the results obtained by applying the statistical methods described in the previous section to StockTwits data are presented. All the analyses have been done in R.

4.1 *Simple correspondence analysis*

The aim of this analysis was to examine whether the aforementioned relationship between the approach and the holding period can be found also among StockTwits' users.

To have a measure of the deviation from the independence of the variables, a χ^2 test was performed on the contingency table that crosses the variables.

Since the χ^2 is not normalized, the Cramer's V (Cramer, 1946) was used to have a clearer measure of the deviation from the independence.

Since the value of Cramer's V depends on the degrees of freedom, i.e., the minimum number between rows and columns, Cohen (2013) defined a rule to interpret the values of the index taking into account this quantity. According to this rule, the thresholds to determine if the connection is strong or weak decreases as the degrees of freedom increase. In this case, Cramer's V is equal to 0.299 and, since the degrees of freedom are four, this value suggests that the connection between *Approach* and *Holding period* is strong, and it is a first confirmation of the presence of the above-discussed relationship.

Moreover, the p-value is < 0.001, therefore the association is statistically significant.

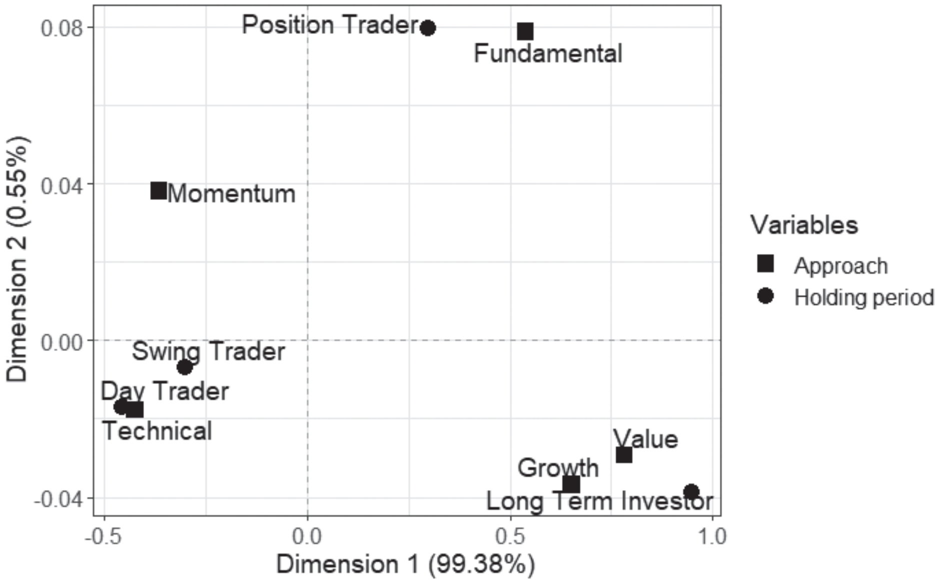

FIGURE 1. - *SCA – Projection of the categories of variables Approach and Holding period on the plane identified by the first two dimensions*

The analysis of eigenvalues shows that the first dimension retains almost all the inertia (99.38%) of the data cloud. This high amount of information explained by the first dimension may be motivated by the strong association between the variables, as evidenced by a Cramer's V value of 0.299, indicating a strong relationship based on Cohen's interpretation. Despite this evidence, to have a better graphical representation, the first two components were chosen to represent the plane reported in Figure 1.

Since the first axis retains the most inertia of the data cloud, the interpretation is based only on it. The quality of representation index (QLT) and the contribution of the categories to the dimensions (CTR) on the first two dimensions are reported in Table 3.

TABLE 3. - *SCA – Quality (QLT) and contribution (CTR) indices*

	Quality		**Contribution**	
Dimension	1	2	1	2
Approach				
Fundamental	0.98	0.02	13.70	52.63
Growth	1.00	0.00	26.61	15.32
Value	1.00	0.00	21.46	5.41
Momentum	0.99	0.01	8.83	17.37
Technical	1.00	0.00	29.40	9.27
Holding Period				
Day Trader	1.00	0.00	17.79	4.55
Swing Trader	1.00	0.00	13.94	1.34
Position Trader	0.93	0.07	5.84	75.27
Long Term Investor	1.00	0.00	62.42	18.84

By jointly observing all the results obtained until now, it can be concluded that the quality of representation of the variables on the dimensions confirms that both the column and row variables are well represented by the first dimension, while the second axis is almost totally not useful to the interpretation.

From the analysis of the biplot in Figure 1, it can be seen that the first component clearly separates investors whose approach is based on technical analysis (*Technical* and *Momentum*) from the investors who base their choices on fundamental analysis (*Value*, *Growth*, *Fundamental*). In addition, the first dimension separates medium-long-term investors (*Position trader* and *Long Term Investor*) from short-term investors (*Swing Trader* and *Day Trader*). Moreover, a gradient from the shorter holding period to the longer one is shown.

Furthermore, generally, weak relationships are identified. The stronger association exists between *Long Term Investor* and *Value*. A weaker association also exists between *Long Term Investor* and *Growth* approach. Since the interpretation is based mainly on the first dimension, it can be noticed also an association between *Long Term Investor* and *Fundamental* approach.

There also exists a repulsive relation between short-term investors and fundamental approaches; while there is an association between short-term investments and technical approaches (*Technical* and *Momentum*). Instead, the opposite is evident for *Position Trader*, indeed, there is an attraction between medium-long-term holding periods and fundamental approaches, while there is repulsion with technical approaches.

These findings are consistent with the meaning of the categories. The association between strategies based on fundamental analysis and long-term investors makes sense because the fundamental analysis is based on companies' performances and macroeconomic factors: all these data need time to be obtained, therefore they cannot be used by a short-term investor. In contrast, the association between short-term investments and strategies based on technical analysis is consistent, because these types of strategies are based on forecasting and historical data, that can be obtained easier and quickly.

4.2 *Multiple correspondence analysis*

The aim of this analysis was to understand if there were some relationships among the experience of the users, the holding period, and their approach.

In order to decide how many dimensions to retain in the analysis, the eigenvalues and the percentage of retained inertia were computed. Nine dimensions were obtained and the first two dimensions were selected using a screeplot. The bidimensional subspace on which the categories of each variable are projected is represented in Figure 2. The first two components retained 30.09% of the total inertia.

In order to improve the interpretation of the results, QLT and CTR indices were computed and reported in Table 4.

The results shown in Figure 2 are coherent with the evidences highlighted by SCA: the first dimension divides the technical approaches from the fundamental ones

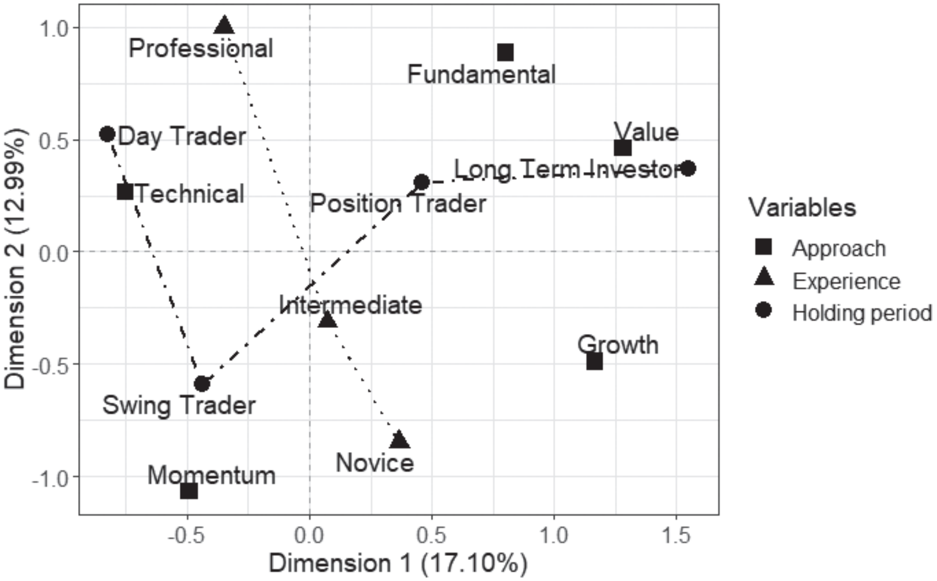

FIGURE 2. - *MCA – Projection of the categories of variables Approach, Holding period, and Experience on the plane identified by the first two dimensions*

and a gradient from the shorter holding period (Day Trader) to the longer one (Long Term Investor) is clearly represented on this dimension.

The associations between categories of the variables *Approach* and *Holding period*, observed in SCA, are maintained in MCA.

TABLE 4. - *MCA – Quality (QLT) and contribution (CTR) indices*

	Quality		Contribution	
Dimension	1	2	1	2
Approach				
Fundamental	0.09	0.11	5.26	8.43
Growth	0.27	0.05	14.78	3.42
Value	0.17	0.02	9.98	1.69
Momentum	0.05	0.24	2.77	17.22
Technical	0.44	0.05	16.05	2.63
Holding period				
Day Trader	0.20	0.08	10.02	5.35
Swing Trader	0.13	0.24	5.12	12.18
Position Trader	0.05	0.02	2.42	1.42
Long Term Investor	0.55	0.03	28.90	2.19
Experience				
Novice	0.04	0.20	1.96	13.61
Intermediate	0.00	0.08	0.15	3.74
Professional	0.06	0.49	2.59	28.10

Regarding the added variable *Experience*, no particular associations are highlighted, but it is interesting to note that the categories *Professional* and *Novice* are well represented on the second dimension and it is interesting to note that, considering the second dimension, a gradient from the most to the less experienced category is shown.

4.3 *Cluster analysis: agglomerative nesting*

In this section, the results of cluster analysis carried out applying the Agglomerative Nesting (AGNES) algorithm to the 23453 StockTwits users are explained. The complete linkage was considered to determine the distance between groups.

The choice of the best number of clusters is based on the indices shown in Table 5, which were computed for different numbers of clusters. Since a large number of clusters, as suggested by Cindex and McClain index, did not provide any insight, the analysis was carried out considering five clusters, as suggested by the other indices. This partitioning is also supported by the stability assessment performed by bootstrapping the Jaccard coefficient.

Since the dendrogram couldn't be plotted due to its size, the clusters were represented using the heatmap shown in Figure 3.

The heatmap shows that AGNES algorithm represents the structure of the Stock-Twtis users on the basis of the variable Approach, indeed, cluster 1 is the cluster of users belonging to the category *Value*, cluster 2 is associated with the category *Technical*, cluster 3 is associated with the category *Growth*, cluster 4 is associated to the category *Fundamental* and, finally, cluster 5 is associated to the category *Momentum*.

This clustering pattern is useful to understand the characteristics of the StockTwits users on the basis of their investment approach: the pattern highlighted by AGNES algorithm shows that users whose investment choices are based on technical analysis (*Technical* and *Momentum*) are characterized by shorter holding periods, while users

TABLE 5. - *Validation indices. The optimal value for each index is highlighted in bold*

Number of clusters	Dunn index	Average silhouette	Cindex	McClain index	Ptbiserial index
2	0.333	0.269	0.311	0.790	0.245
3	0.333	0.336	0.262	0.706	0.380
4	0.333	0.475	0.140	0.529	0.593
5	**0.500**	**0.580**	0.037	0.372	**0.742**
6	0.250	0.526	0.039	0.367	0.732
7	0.250	0.503	0.038	0.363	0.730
8	0.167	0.494	0.040	0.356	0.720
9	0.167	0.484	0.038	0.306	0.690
10	0.200	0.483	0.037	0.301	0.689
11	0.222	0.493	**0.036**	0.296	0.686
12	0.222	0.494	0.038	0.292	0.678
13	0.222	0.498	0.039	0.291	0.676
14	0.222	0.503	0.039	0.286	0.671
15	0.222	0.486	**0.036**	**0.268**	0.666

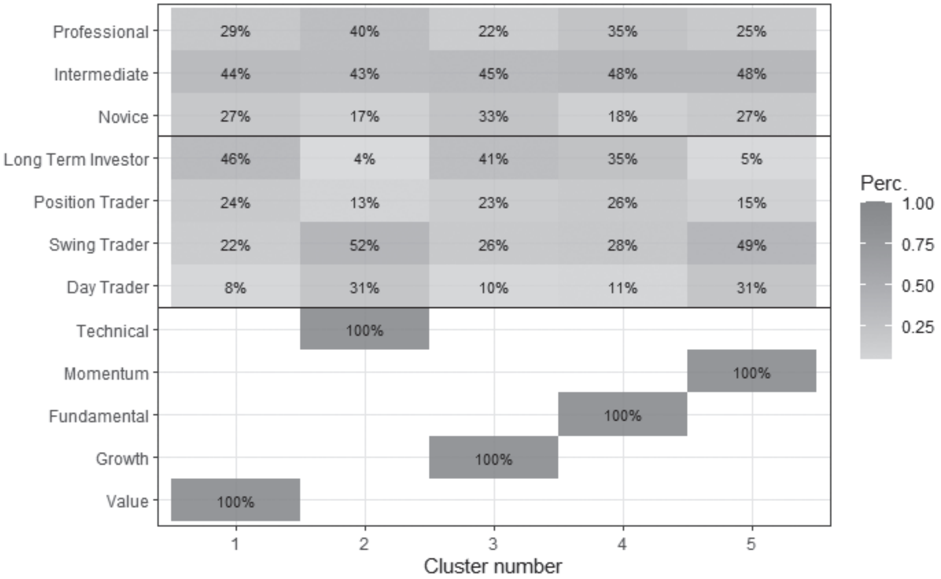

FIGURE 3. - *Heatmap that represents relative distributions of StockTwits users characteristics across clusters. The users are distributed across clusters as follows: Cluster 1 – 2127 users, Cluster 2 – 9940 users, Cluster 3 – 3791 users, Cluster 4 – 2854 users, and Cluster 5 – 4018 users*

whose investment choices are based on fundamental analysis are more medium-long term oriented. Moreover, most of the *Intermediate* and *Professional* users are in cluster 2, the *Technical* cluster.

In this cluster analysis, the users are perfectly classified in each cluster according to their approach. To explain this peculiarity, some considerations about the behaviour of the similarity measures in this specific case must be made. The ordinal variables, as explained by (7), are transformed in such a way as to assume values between 0 and 1, therefore the greater the number of categories, the lower the distance between two subjects characterized by two adjacent categories. Conversely, the distance applied for nominal variables (the *Approach*, here) can be either 0 or 1. Therefore, the nominal variable *Approach* has a greater influence on the distance value and, thus, on the partitioning.

5. CONCLUSIONS AND FUTURE RESEARCH

This work provides an exploration of the self-declared characteristics of StockTwits users. Since users use to fill in their profile with their experience, approach, and primary holding period when trading, on the one hand, it explored the association between these features by means of Correspondence Analysis, and, on the other hand, applied clustering techniques to partition users based on the same traits.

Regarding the former purpose, financial literature showed that a technical strategy in trading leads to short-term investments, whereas a fundamental strategy is strictly connected with medium/long-term investments. Nonetheless, here self-labels were considered, and the same relation was not obvious if considered users self-perception, or, at least, how users want to appear to others. Thus, this study showed that the strength of this relation persists in this context.

Regarding the latter purpose, clustering partitioned users on the basis of their investment approach, detecting groups of technical users preferring shorter holding periods, and fundamental ones declaring to be medium/long-term oriented.

Lastly, although many non-beginners declared to adopt a technical strategy, the association of the experience with the other traits is weak, configuring self-reported experience as a more cross-sectional characteristic.

Further research will explore alternatives to the Gower distance, which equally weighs all types of variables. Additionally, in order to have more insights about this population of users, the authors will leverage the textual self-descriptions, i.e. the *bios*, as an unstructured source of how users show themselves.

ACKNOWLEDGEMENTS

The authors wish to thank the two anonymous reviewers, Professor Marica Manisera and Professor Paola Zuccolotto, whose constructive and valuable comments significantly enhanced the quality of the paper.

REFERENCES

Audrino F., Sigrist F., Ballinari D. (2020). The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting*, **36**(2), 334-357.

Aschenbruck R., Szepannek G. (2020). Cluster validation for mixed-type data. *Archives of Data Science, Series A*, **6**(1), 02.

Benzécri J.P. (1973). *L'analyse des données* (Vol. 2, p. l). Dunod, Paris.

Bolasco S. (2002). *Analisi multidimensionale dei dati: metodi, strategie e criteri d'interpretazione*. Carocci, Roma.

Cazes P. (2014). Simple Correspondence Analysis. In J. Blasius, M. Greenacre (Eds.), *Visualization and Verbalization of Data* (pp. 179-190). Chapman and Hall/CRC, New York.

Cohen J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge, New York.

Cramer H. (1946). *Mathematical methods of statistics*. Princeton University Press, Princeton, NJ.

Dunn J.C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, **4**(1), 95-104.

Everitt B.S., Landau S., Leese M., Stahl D. (2011). *Cluster analysis*. John Wiley and Sons, London.

Gower J.C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857-871.

Graham B., Daniels L. (2015). *The Intelligent Investor*. HarperCollins, New York.

Hennig C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics and Data Analysis*, **52**(1), 258-271.

Husson F., Lê S., Pagès J. (2011). *Exploratory multivariate analysis by example using R* (Vol. 15). CRC press, Boca Raton.

Husson F., Josse J. (2014). Multiple correspondence analysis. In J. Blasius, M. Greenacre (Eds.) *Visualization and Verbalization of Data* (pp. 165-184). Chapman and Hall/CRC, New York.

Jaccard P. (1912). The distribution of the flora in the alpine zone. *New phytologist*, **11**(2), 37-50.

James G., Witten D., Hastie T., Tibshirani R. (2013). *An introduction to statistical learning*. Springer, New York.

Jolliffe I.T. (2002). *Principal component analysis for special types of data*. Springer, New York.

Kamalja K.K., Khangar N.V. (2017). Multiple Correspondence Analysis and its applications. *Electronic Journal of Applied Statistical Analysis*, **10**(2), 432-462.

Kaufman L., Rousseeuw P.J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley and Sons, Hoboken, New Jersey.

Linacre J.M., Rasch G. (2008). The expected value of a point-biserial (or similar) correlation. *Rasch Measurement Transactions*, **22**(1), 1154.

McClain J.O., Rao V.R. (1975). Clustisz: A program to test for the quality of clustering of a set of objects. *Journal of Marketing Research*, 456-460.

Murphy J.J. (1999). *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications*. Penguin Publishing Group, New York.

Oliveira N., Cortez P., Areal N. (2013). On the predictability of stock market behavior using stocktwits sentiment and posting volume. In *Progress in Artificial Intelligence: 16th Portuguese Conference on Artificial Intelligence, EPIA 2013, Angra do Heroísmo,* *Azores, Portugal, September 9-12, 2013. Proceedings 16* (pp. 355-365). Springer, Berlin-Heidelberg.

Oliveira N., Cortez P., Areal N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with applications*, **73**, 125-144.

Pearson K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, **2**(11), 559-572.

Price T.R. (1973). *A Successful Investment Philosophy Based on the Growth Stock Theory of Investing*. books.google.it/books?id=el45HAAACAAJ

Ricciardi R. (2022). What does your self-description reveal about you? A pipeline to analyse StockTwits users. In A. Balzanella, M. Bini, C. Cavicchia, R. Verde (Eds.), *Book of the Short Papers of the 51st Scientific Meeting of the Italian Statistical Society*. (pp. 1809-1814). Pearson, Milano.

Rousseeuw P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, **20**, 53-65.

Sabherwal S., Sarkar S.K., Zhang Y. (2011). Do internet stock message boards influence trading? Evidence from heavily discussed stocks with no fundamental news. *Journal of Business Finance and Accounting*, **38**(9-10), 1209-1237.

Spearman C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology*, (**15**), 201-293.

Taherdoost H., Sahibuddin S., Jalaliyoon N. (2022). Exploratory factor analysis; concepts and theory. *Advances in Applied and Pure Mathematics*, 27, 375-382.

Umar Z., Gubareva M., Yousaf I., Ali S. (2021). A tale of company fundamentals vs sentiment driven pricing: The case of GameStop. *Journal of Behavioral and Experimental Finance*, **30**, 100501.