DATA ANALYTICS 2020

Lost in translation:soft information, sentiment and lending decisions

P. Cerchiello*, S. Filomeni, A. Tanda*, A.M. Toma*

***University of Pavia**
University of Sussex
26th of October
Presenter: Anca Mirela Toma
Email: ancamirela.toma01@universitadipavia.it

Presenter

Toma Anca Mirela is a Research fellow at the University of Pavia, Department of Economics and Management. She is also a PhD fellow in Applied Economics and Management at the University of Bergamo - Pavia. Currently, she works on textual and statistical analysis models applied to fintech risk management. She is also the communication manager for the FinTech-Ho2020 project.

Introduction

- Information is a fundamental component of all financial transactions and markets and it can arrive in multiple forms (Liberti, J.M, Petersen A. M., 2019)
- A large literature has developed around the lending process investigating the role of the information in the credit scoring procedure (Cortes et al. 2019, Campbel et al., 2019, Agrawal et al., 2012)
- The distinction between soft and hard information arose in the finance literature as a way to understand the evolving organization of lenders

Research question

Fact: Human emotions affect corporate financial decisions

Question: Which is the role of sentiment in the lending applications? Therefore:

- HP1a) -> sentiment influences loan decisions
- HP1b) -> sentiment acts as moderator in loan decisions when distance between evaluator and decisor is present.

Methodology

Logistic regression aims at classifying the dependent variable in two groups, characterized by a different status [1=approved vs 0=rejected] in which loans are classified by logistic regression, specified by the following model:

$$ln(\frac{p_i}{1-p_i}) = \alpha + \sum_j \beta_j x_{ij}, \tag{1}$$

It follows that the probability of approval (or rejection) can be obtained as:

$$p_i = \frac{1}{1+exp(\alpha + \sum_j \beta_j x_{ij})}, \tag{2}$$

Sentiment analysis

- Text extraction technique - applied in different context and crucial in our case. Tools: *install.packages("pdftools")*
- Textual analysis procedures for converting qualitative information into quantitative measures
- Score dictionary based: the sentiment score is based on the number of matches between predefined list of positive and negative words and terms contained in each text source (a tweet, a sentence, a whole paragraph);

Data

- Data collecting process involves structured and unstructured information
  - Data used in this study have been manually collected from the credit folders of all (550) mid-corporate loan applications managed (either eventually approved or denied) by the Corporate and Investment Banking Division of a major European bank
  - Time span: from September 2011 to September 2012.
- Data regarding: the borrower, the loan, the lender, distance measures, macroeconomic variables,

Methodology - response-explanatory

The dependent variable *approved* is classified as 1 if the loan has been approved and 0 if has been rejected. The percentage distribution is for status 0 the 13% and the 87% for status 1.

The controls:

- organizational distance measures
- rating measure
- sentiment score
- firm-bank relationship
- loan characteristics
- macroeconomic variables

Methodology - Explanatory variables

Table: Employed Covariates

statistical_rating	Statistical (purely based on quantitative data)
iris_statisticalrating	Statistical+current performance evaluation
integrated_rating	modified statistical rating + qualitative questionnaire (mandatory in the process)
final_rating	integrated rating + potential rating override by the LO
words_clean	Number of words per "commento proposta"
sentiment_score	sentiment measure
dummy_same_branch	dummy equal to 1 if loan officer and loan approver share the same branch
dummy_same_area	dummy equal to 1 if loan officer and loan approver share the same macro area
dummy_same_HQ	dummy equal to 1 if loan officer and loan approver share the same headquarters
log_FD_logit	continuous variable: logarithm of 1 plus the physical distance in kilometers between the branch in which the loan officer responsible for the loan application operates and the bankâs headquarters
log_FD_ols	continuous variable: logarithm of 1 plus the kilometric distance between the branch where the loan officer who conducts the credit scoring operates and the branch of the loan-approving authority
global_guarantee	dummy variable: = 1 if the credit lines of a given borrower are backed by a guarantee of the parent company
collateral	dummy variable: = 1 if the credit line is collateralized
group_belonging	Firmâs belonging to a Group Vs Stand alone company dummy variable = 1 if the borrower is part of an economic group
gradodiindustrializzazioneannual	macroeconomic control ("grado di industrualizzazione according to the macro area of the borrower") - for Italian firms
tassodidisoccupazioneitalia	macroeconomic control ("tasso di disoccupazione") - for Italian firms
italiangdpgrowthannual	macroeconomic control ("crescita annuale GDP - annual %") - for Italian firms
sentperwords_scaled	sentiment score normalized per number of words and scaled
equity_ratio	continuous variable:share of equity over the total assets of the company as stated in the last financial statement
repeated_relationship	dummy variable = 1 if there is a prior lending relationship, =0 if new customer
scope_of_relationship	dummy variable: = 1 if the borrower purchases at least one other banking product from our bank
UD_mlogit	step variable equal to 0 if final rating = integrated; 1 if final rating >integrated rating; 2 if final rating <integrated rating.
dummy_no_override	dummy equal to 1 if rating override is absent height

Textual analysis – I

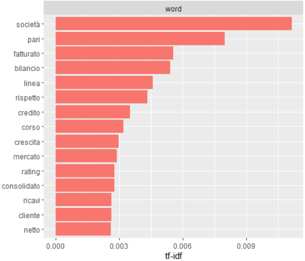

Figure: Left: tf-idf corpus - Right: graphical vizualization

Textual analysis – II

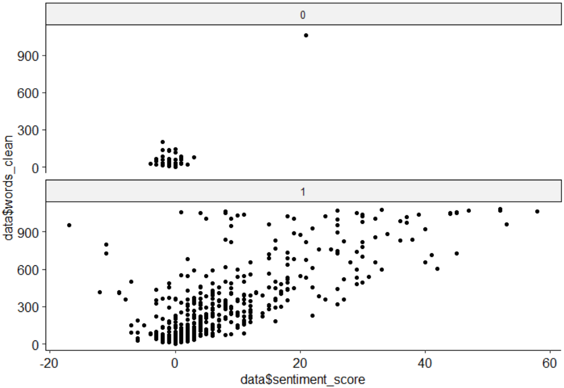

Figure: sentiment and word clean distribution by status

Results – I

Table 4: Logistic regression for different model configurations and interactions

	Dependent variable:					
	approved					
	(1)	(2)	(3)	(4)	(5)	(6)
final_rating	-0.086**	-0.054	0.013	0.157	0.243	
	(0.037)	(0.043)	(0.051)	(0.195)	(0.223)	
dummy_same_branch1	-1.341***	-1.061***	-1.008***	-0.993***	-0.893**	-1.157***
	(0.289)	(0.325)	(0.325)	(0.326)	(0.384)	(0.316)
sentperwords_scaled		0.502***	0.886***	0.887***	0.870***	0.510***
		(0.069)	(0.190)	(0.189)	(0.253)	(0.067)
I(final_rating^2)				-0.008	-0.018	
				(0.010)	(0.012)	
global_guarantee					0.077	
					(0.558)	
collateral					0.579	
					(0.447)	
scope_of_relationship					0.660*	0.771**
					(0.397)	(0.314)
tassodidisoccupazioneitalia					-7.971	
					(5.366)	
final_rating:sentperwords_scaled			-0.039**	-0.040**	-0.025	
			(0.016)	(0.016)	(0.025)	
Constant	3.403***	1.656***	0.946	0.413	0.237	0.881***
	(0.404)	(0.489)	(0.596)	(0.917)	(1.191)	(0.314)
Observations	516	516	516	516	418	550
Log Likelihood	-177.085	-140.220	-137.217	-136.936	-101.769	-149.943
Akaike Inf. Crit.	360.170	288.440	284.434	285.872	223.539	307.885

Results – II

Table 6: Logistic regression for different model configurations and interactions

	Dependent variable:					
	approved					
	(1)	(2)	(3)	(4)	(5)	(6)
dummy_same_branch	-1.163***	-1.061***	-0.790*	-0.807*	-0.941	-1.157***
	(0.314)	(0.325)	(0.459)	(0.463)	(0.630)	(0.316)
final_rating		-0.054	-0.050	-0.062	-0.122**	
		(0.043)	(0.042)	(0.043)	(0.054)	
sentperwords_scaled	0.522***	0.502***	0.578***	0.574***	0.635***	0.510***
	(0.067)	(0.069)	(0.118)	(0.119)	(0.160)	(0.067)
scope_of_relationship				0.889***	0.687*	0.771**
				(0.330)	(0.388)	(0.314)
tassodidisoccupazioneitalia					-6.781	
					(5.054)	
dummy_same_branchsentperwords_scaled			-0.121	-0.131	-0.013	
			(0.146)	(0.146)	(0.202)	
Constant	1.190***	1.656***	1.445***	1.188**	1.930**	0.881***
	(0.292)	(0.489)	(0.548)	(0.566)	(0.827)	(0.314)
Observations	550	516	516	516	418	550
Log Likelihood	-153.071	-140.220	-139.868	-136.094	-104.463	-149.943
Akaike Inf. Crit.	312.141	288.440	289.737	284.189	222.925	307.885

Note: *p<0.1; **p<0.05; ***p<0.01

Preliminary conclusions and ongoing research

- Sentiment can impact lending decisions together with variables related to the distances that captures well the informative disclosure issues and communication biases in the borrower - lender decision process.
- It can act as a predictor on the probability of approval or rejection
- It can act as a moderator in the channelling of soft/privileged information
- but is also able to predict default on loans -> ongoing focus on default target variable & credit granted

Thank you for you attention!