Report: First Draft

E404 - Natural Language Processing
Autumn 2023

Professor: Osama Khalid

Professor Assistant: Tomas Pacheco

Student: Alfonso Sundblad

Abstract

This research investigates the linguistic dimensions of economic uncertainty, with a specific focus on the Spanish-speaking community in Argentina. We explore linguistic patterns across various sources, spanning online text, political discourse, and spoken news transcriptions. By examining linguistic shifts during periods of economic instability, we aim to provide valuable insights into how language reflects and responds to economic uncertainty.

The novelty of this study lies in its pioneering exploration of linguistic patterns in the context of economic uncertainty in the Spanish language, an area that has received limited attention. Through a combination of natural language processing tools and techniques inspired by related literature, we aim to provide fresh insights into the correlation between linguistic style and economic conditions, shedding light on how public discourse is molded by economic uncertainty, and highlighting the potential significance of linguistic analysis in understanding socioeconomic dynamics.

Introduction

Originally my final project I had a different original theme, which initially addressed "Economic Uncertainty and Mental Health" with a Naive Bayes Classifier. However the strength of this was limited. As we couldnt directly analyze mental health we were restricted to the linguistic style of it or a sentiment analysis, in which the new alternative was far more interesting.Thus, we shifted our focus slightly, with a similar question at heart but much more interesting: "The Linguistic Style of Economic Uncertainty." This adaptation allows me to analyze people"s language in a continuous and more detailed manner, rather than using a more specific classifier-based approach.

In other words, we will be focusing in studying the linguistic style associated with economic uncertainty. For this we will analyse linguistics of different types of sources, from online written text, political publicity and speeches, to spoken news transcriptions. On the other hand, for economic uncertainty we will start using a standard risk index calculated by JP Morgan: EMBI (and potentially some of its variations). We acknowledge the limitation and bias of our sources selection, however we think including different types, styles of speech and different society groups empling it we can help to mitigate it. However, once the methodology for this analysis is set up, we can expand even further the information reach. Right now in this first draft, we are limiting ourselves to just posts and comments on specific Argentine subreddits.

It is essential to emphasize that this type of study is relatively underexplored in the Spanish language, which underscores the novelty and significance of this research. To thoroughly investigate this subject, I intend to employ various tools of linguistic analysis, such as Linguistic Inquiry and Word Count (LIWC), different readability assessments, parts-of-speech analysis, and character-level feature examination. However once again, for this first draft we will be limiting ourselves to Spanish LIWC.

The primary hypothesis I am making is that there will be observable shifts in the linguistic style of discussions related to economic uncertainty over time. I expect that during periods of economic instability, there will be an increase in negative sentiment and the usage of specific economic terms related to uncertainty. Conversely, during periods of economic stability,

I anticipate a more positive sentiment and a different set of terms used in discussions. In a second stage, this research seeks to examine the potential relationship with depression like language.

Related Works

To start with, some of the related works and papers we will be mentioning as reference are not specifically about Economics but rather Computer Sciences, Political Sciences, etc. This is due to the novelty of these tools in economics research.

In the first place, the study by Aromí (2022), while being a working paper, inspired me to use these technologies in economic problems. It also gives us the possibility of creating an uncertainty index to use in addition to the country risk. "This work proposes an index that describes the economic opinions conveyed by Argentine users on the social network Twitter. After identifying economic messages, these are classified based on the frequency with which words associated with uncertainty are used. Qualitative evaluation of the index suggests a strong link with relevant economic and political events. Estimations of statistical models indicate that the index contains information about the economic cycle, consumer confidence, and the evolution of the foreign exchange market. Supplementary analyses demonstrate that the focus on the concept of uncertainty and the use of natural language processing techniques constitute key elements for the satisfactory performance of this opinion indicator."

Furthermore, the findings of Vandoros et al. (2019) and Vandoros & Kawachi (2021) provide a fundamental perspective on suicide/mental health and periods of economic uncertainty. Although they do not use a methodology similar to that of my thesis, they provide a justification for this research. I consider them foundational literature from which I depart with new methodologies to investigate a similar theme.

In this recently published paper (Charemza et al., 2022), NLP is used to delve into measuring economic and political uncertainty. It is useful to see how they process and scrape data from Russian newspaper articles. While the ultimate goal is not so similar, the tools and methods used serve as a guide for several elements already mentioned in the thesis, such as the use of LIWC and both filtering, refining, and scraping virtual newspapers. It also shows the

potential that these tools can have for analyzing countries with data that is not always available or reliable. "The article proposes a method for constructing country-specific text-based measures of economic policy uncertainty. To avoid translation issues and human validation costs, we apply natural language processing and sentiment analysis to build such measures for Russia. We compare our measure with the one previously developed through direct translations from English and human validation. In this comparison, our measure performs equally well in assessing uncertainty related to key events that affected Russia between 1994 and 2018 and performs better in detecting the effects of uncertainty on Russia"s industrial production."

Additionally, Nanath et al. (2022) proved to be helpful and an example to explore a possible way to relate depression-like language to the results of the first half of the model. They provide an extremely useful framework for this type of analysis, with the only caveat being that it is in English. "Governments worldwide have implemented rigorous restrictions to curb the spread of the COVID-19 pandemic. While beneficial for physical health, these preventive measures could have a profound adverse effect on the mental health of the population. This study focuses on the impact of lockdowns and mobility restrictions on mental health during the COVID-19 pandemic. First, we developed a novel mental health index based on the analysis of data from over three million tweets worldwide, using the machine learning approach of Microsoft Azure. Then, the calculated mental health index scores are regressed against the lockdown stringency index and the Google mobility index using ordinary least squares (OLS) regression with fixed effects. The results reveal that reduced labor mobility, reduced mobility in stores and recreational activities, and increased residential mobility (confinement at home) have negatively affected mental health. However, mobility restrictions in parks, grocery stores, and pharmacies were found to have no significant impact. The proposed mental health index provides a path for theoretical and empirical studies of mental health using social media."

The decisions in Penczynski (2019) for classifying communications are of great help in making decisions about the model to be developed. Fine-tuning can sometimes be arbitrary and an extremely tedious process, which makes this paper useful for saving time and justifying fine-tuning decisions.

We know that using processed text as data can be risky. Like any tool, it has its strengths and weaknesses, as well as potential pitfalls in its use. Benoit et al. (2009) wrote a paper in detail about the possible problems, biases, and errors of these tools, which allows us to either strengthen or accept them.

On the other hand, the work of Yang & Srinivasan (2016) is an excellent example of methodology and the use of social networks, in their case, Twitter, for the analysis of people"s daily life behavior. "Life satisfaction refers to a relatively stable cognitive assessment of one"s own life. Life satisfaction is an important component of subjective well-being, the scientific term for happiness. The other component is affect: the balance between the presence of positive and negative emotions in daily life. While affect has been studied using social media data sets (particularly Twitter), life satisfaction has received little or no attention. In this study, we examine trends in posts about life satisfaction from a two-year Twitter data sample. We apply a surveillance methodology to extract expressions of both satisfaction and dissatisfaction with life. One notable result is that, according to their definitions, trends in posts about life satisfaction are immune to external events (political, seasonal, etc.), unlike trends in affect reported by previous researchers. When comparing users, we find differences between satisfied and dissatisfied users in various linguistic, psychosocial, and other aspects. For example, the latter post more tweets expressing anger, anxiety, depression, sadness, and about death. We also study users who change their status over time, going from being satisfied with life to dissatisfied or vice versa. It is noteworthy that the characteristics of the psychosocial tweets of users who change from satisfied to dissatisfied differ significantly from those who remain satisfied over time. In general, the observations we made are consistent with intuition and with observations in social sciences research. This research contributes to the study of people"s subjective well-being through social networks."

(Ramírez-Esparza et al., 2007) once again addresses this topic, but from a perspective that is extremely useful to us: in Spanish. It is not a particularly detailed paper on a specific analysis objective, but it develops and explains NLP tools for psychological analysis in Spanish. "The Linguistic Inquiry and Word Count (LIWC, for its acronym in English, Pennebaker, Francis, & Booth, 2001) is a computer program that analyzes texts. This program calculates the

percentage of words within a text according to several dozen categories. The reliability of this program has been widely demonstrated in the English language. In this research, two studies were conducted to analyze the equivalence of the LIWC program in Spanish to the program in English. Study 1 presents the procedure for translating LIWC from English to Spanish and demonstrates the equivalence between LIWC categories in English and their corresponding categories in the Spanish version. Study 2 shows the use of LIWC in both English and Spanish by comparing the language used by women in depression and breast cancer discussion forums on the Internet. The results showed that the versions correlate in most categories. It was also found that women in depression forums use different word categories than women in breast cancer forums, and these differences are similar in discussion forums in both Spanish and English. The implications of using this program within the Spanish language are discussed."

Like the previous paper, Pérez et al. (2021) presents a sentiment classification tool in Spanish, but the particularity is that it is in Argentine Spanish. "The extraction of opinions from texts has generated great interest in recent years, as we are experiencing an unprecedented volume of user-generated content on social networks and other places. A problem that social researchers encounter when using opinion mining tools is that they are often behind commercial APIs and are not available for languages other than English. To address these problems, we present pysentimiento, a multilingual Python tool for Sentiment Analysis and other Natural Language Processing (NLP) tasks in social environments. This open-source library offers state-of-the-art models for Spanish and English seamlessly, allowing researchers easy access to these techniques."

Finally, Khalid & Srinivasan (2020) provide a solid application and a variety of tools for the analysis of linguistic styles in: social networks like 4chan and Reddit; and accumulated news with Voat. “Historically, content has been the main focus for studying language in online communities. Instead, this article focuses on the linguistic style of communities. While we know that people have distinguishable styles, we ask whether communities have distinguishable styles. Furthermore, whereas previous work has relied on a narrow definition of style, we employ a broad definition involving 262 features to analyze the linguistic style of 9 online communities on 3 social media platforms discussing politics, television, and travel. We found that communities

indeed have distinctive styles. Furthermore, style is an excellent predictor of group membership (F-score of 0.952 and accuracy of 96.09%). Although, on average, it is statistically equivalent to predictions using only content, it is more robust to reductions in training data"

Methods

The range of data we will be studying spans from 1 January 2018 to 31 December 2022. This timeframe encompasses a period of significant economic events and fluctuations, allowing for a comprehensive analysis of language responses to economic uncertainty in the Spanish-speaking community of Argentina. Including a change in presidency and 2 years prior and after COVID-19. In Table 3, note that there are some channels highlighted in red, this means they are not going to be included since they do not fit the time period. Additionally at the time being we were limited in the extraction of the re highlighted subreddits, and if we would want to pursue their analysis would have to change the method of retrieving that type of information.

To start with, as previously mentioned, in order to answer our research question, we will be using the risk index EMBI to represent the economic uncertainty in Argentina. The historical data of this index was hard to retrieve and we relied on the daily recompilation of the Central Bank of Chile of this value for Argentina. However it has about 20% of days missing in our time frame, so we are working towards either calculating the missing data or retrieving it from another source (still the same index).

For the linguistics side of our evaluation we have several sources, to start with: Reddit. We seek to analyze the top 5 Argentine subreddits (Table 1). The data will be extracted from the-eye organization, a mirror of the pushshift archive. It includes the title, body of the post and their comments. Unfortunately we were only able to get 3 of the 6 subreddits we originally seeked. The ones missing at the time, as previously mentioned, are highlighted in red. Also is important to note that posts that just contain images, links or have either title or body not accessible will be completely scrapped, as we join these into one datapoint (keeping comments as a separated source)

For Online Newspapers the 6 most read papers are being analyzed, where two are more specific towards economics. These are stated in Table 2. These are according to a research firm,

SimilarSites, that generated a ranking with traffic, engagement, and a ratio of news returning users (Figure 1). These will be extracted with web scraping using: Scrapy, Selenium and BeautifulSoup. Specifically we will be retrieving headlines and complete articles within the time frame.

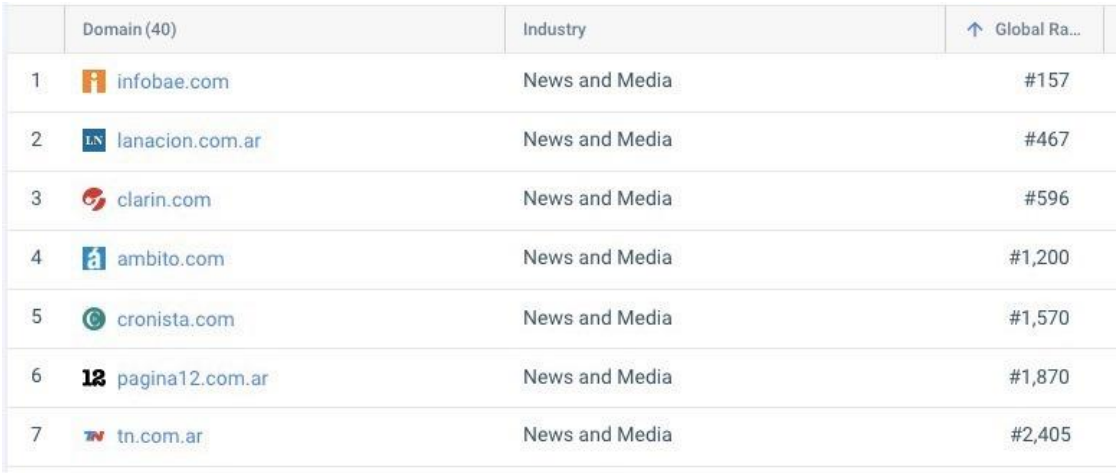
For YouTube we have chosen both national and provincial level official channels, as well as an official Tv channel, and the 5 popular Tv news channels (Table 3.). Once again, as previously stated unusable channels are highlighted in red. We used video titles and transcripts of the complete video as datapoints. For these, a headless python API will be used which allows you to get the transcripts/subtitles for a given YouTube video, which also works for automatically generated subtitles(https://pypi.org/project/youtube-transcript-api/). In addition to this we will use *scapetube*(https://github.com/dermasmid/scrapetube) to get all the IDs from the channel and *pytube*(https://pytube.io/) to get the precise date and title of each video.

Finally, we apply Spanish LIWC to these now processed texts from the several different sources. Now we have to compute the correlations between the daily risk index EMBI and each of the LIWC parameters (word count, function words, total pronouns, etc.). However for each value of EMBI we have several results of our Spanish LIWC values within each source. To remedy this we group the values by date and calculate an average.

An additional note is that currently we are only using Spanish LIWC, but we wish to expand our combination of natural language processing techniques, inspired from Khalid, O., & Srinivasan, P. (2020). “Style Matters! Investigating Linguistic Style in Online Communities”: Readability(ARI, CLI, etc.), Parts of Speech Frequencies(verbs, nouns, etc.), Character Frequencies(white space, punctuations, etc.) and Word level features(Spanish LIWC, Simpson"s Index, etc.)

AUXILIAR RESOURCES:

FIGURE 1.



	Domain (40)	Industry	Global Ra...
1	infobae.com	News and Media	#157
2	lanacion.com.ar	News and Media	#467
3	clarin.com	News and Media	#596
4	ambito.com	News and Media	#1,200
5	cronista.com	News and Media	#1,570
6	pagina12.com.ar	News and Media	#1,870
7	tn.com.ar	News and Media	#2,405

TABLE 1.

Sub Reddits	Url
r/argentina	https://www.reddit.com/r/argentina/
r/RepublicaArgentina	https://www.reddit.com/r/RepublicaArgentina/
r/Republica_Argentina	https://www.reddit.com/r/Republica_Argentina/
r/AskArgentina	https://www.reddit.com/r/AskArgentina/
r/ArgenGaming	https://www.reddit.com/r/ArgenGaming/
r/dankgentina	https://www.reddit.com/r/dankgentina/

TABLE 2.

News Papers	Url
Infobae	https://www.infobae.com
La Nacion	https://www.lanacion.com.ar
Clarin	https://www.clarin.com
Pagina 12	https://www.pagina12.com.ar
Ambito Financiero	https://www.ambito.com
El Cronista	https://www.cronista.com

TABLE 3.

Youtube Channels	Category	URL
Casa Rosada	Republic Government Argentina	https://www.youtube.com/@casarosada
Senado	Republic Government Argentina	https://www.youtube.com/@SenadoTVArgentina
Diputados	Republic Government Argentina	https://www.youtube.com/@diputados.argentina
Poder Judicial	Republic Government Argentina	https://www.youtube.com/@pjn-videoconferencias
Ministerio de Ambiente y Desarrollo Sostenible	Government Ministries	https://www.youtube.com/@ministeriodeambienteydesar2599
Ministerio de Ciencia, Tecnología e Innovación	Government Ministries	https://www.youtube.com/@MinisterioDeCiencia
Ministerio de Cultura	Government Ministries	https://www.youtube.com/@culturaargentina
Ministerio de defensa	Government Ministries	https://www.youtube.com/@Mindefargentina
Ministerio de desarrollo Social	Government Ministries	https://www.youtube.com/@MDSNacion
Ministerio de desarrollo Territorial y Hábitat	Government Ministries	https://www.youtube.com/@minhabitat
Ministerio de Economía	Government Ministries	https://www.youtube.com/@MinisteriodeEconomiaAR
Ministerio de Educación	Government Ministries	https://www.youtube.com/@educacionar
Ministerio del Interior	Government Ministries	https://www.youtube.com/@MinisteriodelInteriorAR
Ministerio de Justicia y derechos Humanos	Government Ministries	https://www.youtube.com/@MinisteriodeJusticiayDDHH
Ministerio de Las Mujeres, Géneros y Diversidad	Government Ministries	https://www.youtube.com/@MinGenerosAR
Ministerio de Obras Públicas	Government Ministries	https://www.youtube.com/@ObrasPublicasAr
Ministerio de Relaciones Exteriores, Comercio Internacional y Culto	Government Ministries	https://www.youtube.com/@CancilleriaARG
Ministerio de Salud	Government Ministries	https://www.youtube.com/@MsalNacion
Ministerio de Seguridad	Government Ministries	https://www.youtube.com/@minseg
Ministerio de Trabajo, Empleo y Seguridad Social	Government Ministries	https://www.youtube.com/@mintrabajo
Ministerio de Transporte	Government Ministries	https://www.youtube.com/@MinisteriodeTransporteArg
Ministerio de Turismo y deportes	Government Ministries	https://www.youtube.com/@TurismoNacionAR
Provincia Buenos Aires	Provincial Government	https://www.youtube.com/@BAProvincia
Ciudad Autónoma de Buenos Aires	Provincial Government	https://www.youtube.com/@GCBA
Provincia Catamarca	Provincial Government	https://www.youtube.com/@GobiernodeCatamarca
Provincia Chaco	Provincial Government	https://www.youtube.com/@ChacoGobiernodeTodos

Provincia Chubut	Provincial Government	https://www.youtube.com/@gobiernochubut3215
Provincia Córdoba	Provincial Government	https://www.youtube.com/@GobCordobaok
Provincia Corrientes	Provincial Government	https://www.youtube.com/@GobiernodeCorrientess
Provincia Entre Ríos	Provincial Government	https://www.youtube.com/@GobiernoER
Provincia Formosa	Provincial Government	https://www.youtube.com/@GobiernodeFormosaOficial
Provincia Jujuy	Provincial Government	https://www.youtube.com/@GobiernodeJujuyOK
Provincia La Pampa	Provincial Government	Doesnt Have a Youtube Channel
Provincia La Rioja	Provincial Government	https://www.youtube.com/@GobiernoDeLaRiojaES
Provincia Mendoza	Provincial Government	https://www.youtube.com/@mendozagob
Provincia Misiones	Provincial Government	https://www.youtube.com/channel/UCkvikZrVh7Gp1yLAWUYPLng
Provincia Neuquén	Provincial Government	https://www.youtube.com/@NeuquenProvincia
Provincia Río Negro	Provincial Government	https://www.youtube.com/@rionegrogob
Provincia Salta	Provincial Government	https://www.youtube.com/@GobiernodeSaltaOficial
Provincia San Juan	Provincial Government	https://www.youtube.com/@sisanjuanok
Provincia San Luis	Provincial Government	https://www.youtube.com/@sanluisgobierno
Provincia Santa Cruz	Provincial Government	https://www.youtube.com/@santacruzar1495
Provincia Santa Fe	Provincial Government	https://www.youtube.com/@gobsantafe
Provincia Santiago del Estero	Provincial Government	https://www.youtube.com/@gobiernodesantiagodelestero
Provincia Tierra del Fuego	Provincial Government	https://www.youtube.com/@gobiernotierradelfuego1540
Provincia Tucumán	Provincial Government	https://www.youtube.com/@prensatucuman
TV Publica	National TV Channel	https://www.youtube.com/@TVPublicaArgentina
TN	News Channel	https://www.youtube.com/@todonoticias
La nacion +	News Channel	https://www.youtube.com/@lanacion
Cronica	News Channel	https://www.youtube.com/@cronicatv
A24	News Channel	https://www.youtube.com/@A24com
C5N	News Channel	https://www.youtube.com/@c5n

Bibliography

- Aromí, J. D. (2022). Medición de Incertidumbre Económica en Redes Sociales en Base a Modelos de Procesamiento de Lenguaje Natural. *Working Papers*. https://ideas.repec.org/p/aoz/wpaper/179.html

- Benoit, K., Laver, M., & Mikhaylov, S. (2009). Treating words as data with error: Uncertainty in text statements of policy positions. *American Journal of Political Science*, *53*(2). https://doi.org/10.1111/j.1540-5907.2009.00383.x

- Charemza, W., Makarova, S., & Rybiński, K. (2022). Economic uncertainty and natural language processing; The case of Russia. *Economic Analysis and Policy*, *73*, 546–562. https://doi.org/10.1016/J.EAP.2021.11.011

- Khalid, O., & Srinivasan, P. (2020). Style Matters! Investigating Linguistic Style in Online Communities. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 360–369. https://doi.org/10.1609/ICWSM.V14I1.7306

- Nanath, K., Balasubramanian, S., Shukla, V., Islam, N., & Kaitheri, S. (2022). Developing a mental health index using a machine learning approach: Assessing the impact of mobility and lockdown during the COVID-19 pandemic. *Technological Forecasting and Social Change*, *178*. https://doi.org/10.1016/j.techfore.2022.121560

- Penczynski, S. P. (2019). Using machine learning for communication classification. *Experimental Economics*, *22*(4). https://doi.org/10.1007/s10683-018-09600-z

- Pérez, J. M., Giudici, J. C., & Luque, F. (2021). *pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks*. https://arxiv.org/abs/2106.09462v1

- Ramírez-Esparza, N., Pennebaker, J. W., García, F. A., & Suriá, R. (2007). The psychology of word use: A computer program that analyzes texts in Spanish. *Revista Mexicana de Psicología*, *24*(1), 85–99. https://psycnet.apa.org/record/2007-08364-008

- Tham, W. W., Sojli, E., Bryant, R., & McAleer, M. (2021). Common Mental Disorders and Economic Uncertainty: Evidence from the COVID-19 Pandemic in the U.S. *PLOS ONE*, *16*(12), e0260726. https://doi.org/10.1371/JOURNAL.PONE.0260726

- Vandoros, S., Avendano, M., & Kawachi, I. (2019). The association between economic uncertainty and suicide in the short-run. *Social Science & Medicine*, *220*, 403–410. https://doi.org/10.1016/J.SOCSCIMED.2018.11.035

- Vandoros, S., & Kawachi, I. (2021). Economic uncertainty and suicide in the United States. *European Journal of Epidemiology*, *36*(6), 641–647. https://doi.org/10.1007/S10654-021-00770-4/TABLES/3

- Yang, C., & Srinivasan, P. (2016). Life Satisfaction and the Pursuit of Happiness on Twitter. *PLOS ONE*, *11*(3), e0150881. https://doi.org/10.1371/journal.pone.0150881
- Youtube-Transcript-Api. https://pypi.org/project/youtube-transcript-api/.
- ScapeTube. https://github.com/dermasmid/scrapetube
- Pytube. https://pytube.io/
- Central Bank of Chile. (2023). Spreads soberanos EMBI Global Spread (promedio, puntos base) https://si3.bcentral.cl/siete/ES/Siete/Cuadro/CAP_EI/MN_EI11/EI_EMBI?cbFechaDiaria=2018&cbFrecuencia=BUSINESS&cbCalculo=NONE&cbFechaBase= Descripción
- SimilarWeb. (2023). Website Ranking, filter to Argentine online newspapers. https://pro.similarweb.com