

IFC-Bank of Italy Workshop on “Machine learning in central banking”

19-22 October 2021, Rome / virtual event

Using twitter data to measure inflation perception¹

Julien Denes, Ariane Lestrade and Lou Richardet,
Bank of France

¹ This presentation was prepared for the Workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the central banks and other institutions represented at the event.

Using Twitter Data to Measure Inflation Perception

A working paper

Julien Denes

Ariane Lestrade

Lou Richardet

Banque de France

Abstract

Anchoring inflation expectations and measuring the current and future impact of prices evolution is a crucial issue for central banks. With the current rise of social networks, a new source of information has appeared to measure inflation perception. In this study, we propose an indicator of inflation perception based on Twitter data, focusing on the specific community of users who retweeted posts of the Banque de France account. Despite the bias induced, this strategy allows to efficiently extract information from an expert community on economic and financial subjects, capturing a potentially more relevant signal. We combine supervised machine learning and natural language processing methods with dictionary-based filters to classify all tweets posted by these retweeters, by first detecting whether they relate to prices issues, and then assessing which of the selected tweets mention inflation, deflation or is off-topic. Finally, we create a Twitter indicator of inflation perception, which is the difference between the number of tweets about inflation and the number of tweets about deflation. The resulting Twitter indicator is consistent with monthly household surveys on inflation expectations and perception, and is highly correlated with the inflation rate. We also show that it has a strong anticipatory power of the future inflation. These results suggest that it is both possible and relevant to use Twitter data to construct a daily measure of inflation perceptions.

JEL Classification: E31, C53, C55, D84, E58.

Keywords: inflation, inflation perceptions, inflation expectations, Twitter data, text mining, machine learning, big data, real-time data, high-frequency data.

This working paper describes preliminary results of ongoing research. It is made available to the public solely to elicit discussion and comments. Views expressed in the paper are those of the authors and do not reflect the position of the Banque de France.

Authors' email address: julien.denes@banque-france.fr.

1. Introduction

Maintaining price stability over the medium term is one of the most essential task of central banks as it highly influences the trust economic agents have on the future. This raises the key question of how to measure inflation, but most importantly how to measure the inflation expectations of the population in order to anticipate their behavior. Formally, inflation is defined by the French National Institute for Statistics (INSEE) as the loss of the purchasing power of money, which translates into a general and sustainable increase in prices.¹ To measure this phenomenon on the long run, the Harmonized Index of Consumer Prices (HICP) is the most used tool, in particular because it was designed to allow for international comparison. Calculated monthly, it makes it possible to estimate, between two given periods, the average change in the prices of products consumed by households. The methodology is harmonized within the European Union to allow for comparisons between each member state national Consumer Prices Index (CPI).

The most crucial question however is rather how economic actors perceive, anticipate, and react to inflation. To measure this perception of inflation, several sources can be used. First, some indicators are constructed using surveys. For instance, the European Central Bank (ECB) produces the Survey of Professional Forecasters, who are asked to forecast inflation rate among many other macroeconomic values. Likewise, the Consensus Economics survey and the Eurozone Barometer are both published every month. Some surveys focus specifically on inflation, such as the Atlanta Fed Business Inflation Expectations, in which participants provide their estimations of future inflation as values, or assign probabilities that inflation is within predefined ranges. Some other indicators rather survey the perception of non-professional individuals within the general population. For instance, INSEE's Monthly Household Survey includes a few questions about perceived past inflation and expected future inflation rate.

Second, indicators can also be extracted from financial markets, for instance using "break-even" inflation rates and inflation swap markets. Break-even inflation rate is the difference between the yield of a nominal bond and an inflation-linked bond of the same maturity. A swap is a product that converts an inflation-indexed loan (or borrowing) into a fixed-rate loan (or borrowing). Unlike surveys, these indicators are available at a high frequency and react more quickly about changes in the economy.

However, both those sets of indicators have some limits. For instance, if respondents to the ECB Survey of Professional Forecasters are convinced of the Bank's credibility, their answers will reflect this opinion rather than their true inflation expectations. Market indicators also have their limits, since observed prices incorporate risk and liquidity premia and may carry a seasonality bias, which can be problematic for extracting information on inflation expectations. The best way to obtain an unbiased measure of inflation is therefore to have as much indicators as possible, taking into account the limitations of each of them.

It is in this perspective that this study is inscribed. Our goal is to provide an alternative measure of inflation perception, by exploiting information made available by social networks. The current boom in social networks indeed provides a new source of information to find out how individuals feel about price trends. Twitter in particular allows relatively open access to its massive data, with millions of tweets being published each month, just in French. In order to keep the amount of data collected and analyzed within a reasonable range, this study focuses on a specific subset of users, namely the retweeters of the Banque de France Tweeter account. This community has been the subject of a previous internal study of Banque

¹ <https://www.insee.fr/fr/metadonnees/definition/c1473>

de France (Kintzler, 2018), which highlights among others the following characteristics: most retweeters are French, located in Paris and work in the banking, finance or economic sector. They seem therefore to be well informed about price evolution than the general population, and make them a population of interest to survey. Some bias may of course arise in comparison with data obtained from the general population, but we hypothesize that the opinion of informed experts on the topic are of higher value to estimate the general perception of inflation.

Recent studies have explored the contribution of indicators that measure sentiment perception about macroeconomic issues based on social media or newspaper data. For instance, Bertoli, Combes and Renault (2017) have calculated a media sentiment indicator for short-term employment forecasting. Thorsrud (2016) uses the content published by some Norwegian media to obtain a leading indicator of activity in Norway. Finally, Baker, Bloom and Davis (2016) create an economic policy uncertainty (EPU) index based on newspaper coverage frequency and demonstrate it proxies well for movements in policy-related economic uncertainty. Altig *et al.* (2020) then generalize it by successfully applying their methodology to Twitter, creating a high frequency uncertainty index. Despite this quite rich literature, very few studies focus on analyzing social media data regarding inflation problematics.

Angelico *et al.* (2021) from Banca d'Italia published one of the pioneering studies in the use of Twitter data for measuring inflation perception. The authors first collect all tweets in Italian related to prices, and then filtered appropriate ones using topic models to reduce noise. They then identify tweets related to increase of inflation and decrease of inflation using keywords, and finally combine create an indicator of perceived inflation.

Inspired by these results, we propose a novel approach than also combines the two approaches, namely keywords and natural language processing methods. Our final objective is to improve the performance of the final indicator, in particular by using modern state-of-the art supervised machine learning techniques rather than unsupervised topic models. Our methodology is conducted in three steps, focusing on the case of France. First, we collect all tweets published by all accounts that have retweeted a tweet from the Banque de France account. Second, we keep only those related to prices by classifying them using word2vec embeddings and random forests. Third, we classify each tweet in one of three categories: inflation, deflation, or other. Finally, we construct our indicator as the difference in the number of tweets related to inflation minus those related to deflation. We finally show that our indicator is highly correlated to more traditional survey-based metrics for the perception of inflation.

The remainder of the paper is structured as follows. We first describe the Twitter data used to create the indicator of price perception. Then, we will explain our methodology and its consecutive steps. Finally, we present the results and show their predictive power of the general perception of inflation.

2. Twitter data

In this study, we use Twitter data, which are increasingly used in economic news reporting. Twitter is a social network where users can publish short posts called tweets. Each day, millions of tweets are posted in French. From preliminary experiments, we even measured that thousands of tweets are posted each day that mention the simple keyword “prix” (which translates to both “price” and “prices” in French).

For the purpose of this study, in order to keep the amount of tweets reasonable, this study focuses on a specific subset of users, namely those who retweeted at least one post of the official Banque de France Twitter account. A previous internal research paper from Banque de France (Kintzler, 2018) collected this list of users and analyzed its population. It appears to be mostly French, located in Paris,

and working in the banking, finance or economic sector, as almost half of them have keywords related to finance or economy in their account description. Appendix 7.4.3 presents the methodology used for this analysis. Even though these users are not representative of the general population, extracting signals from their tweets could be insightful because of their interest in economic matters. In future works, we will endeavor to generalize our methodology to a much larger and more representative subset of Twitter users, and if allowed to the whole population.

In order to not be limited by the restricted number of tweets imposed by the official Twitter API, we used the open-source tool Twint² to construct our database. Using our list of Banque de France retweeters, we collected for each of them the whole history of their public post, referred to as “timelines”, with the only condition that the detected language of the tweet must be French. In total, the list of retweeters is composed of 3,548 accounts. We collected tweets from January 2008 to June 2021, although further filtering is applied on the following steps. In total, the number of all posts in all 3,548 timelines amounts to 10,382,847 tweets.

The information available at the tweet level is very rich: the text of the tweet, the date of creation, the geolocation, the language (in our case only French), the number of times it was shared or bookmarked, etc. A twitter account is also characterized by several variables: the date of creation of the profile, the short biography provided by the user, the number of accounts followed, the number of followers, and so on.

3. Methodology

The biggest challenge of this study relies on succeeding to detect the tweets related to prices among all possible topics discussed. Preliminary experiments showed us that using only is insufficient, because the targeted theme is not precise enough to avoid the problematic of linguistic polysemy of the French language. A very concrete example is the word “prix”, which means both “price” (both singular and plural), and “award” or “prize”. Therefore, using a list of keywords that is too large will lead to selecting too many tweets not related to inflation (“false positives”). On the other hand, a very restrictive list of keywords may lead to filter out many relevant posts (“false negatives”). Using machine learning, which is a much more flexible and precise tool, arises as a promising alternative. However, filtering the whole database of Tweets could be much too expensive in terms of computing power and time, since machine learning models rely on complex calculation, whereas keywords identification are extremely simple and fast.

To leverage both tools as efficiently and as precisely as possible, we use multi-level filtering, in which we combine keywords filters and machine learning filters to detect tweets relevant for the analysis. The methodology consists of four steps filtering and classifying tweets. First, a dictionary-based filter is applied to keep only tweets containing specific keywords related to the topic of prices. Second, a supervised Machine Learning model is trained on a random sample of 800 manually labelled tweets, and is used to the rest of the tweets in order to clean the residual noise of the first filter and retaining only tweets related to prices matter. Third, a keywords-based method is used to classify the direction of prices evolution mentioned in the tweets between “inflation”, “deflation”, or “other”. Fourth, tweets mentioning foreign prices are excluded from the analysis since the study focuses on French prices.

The following sections detail each of these four steps, and present evaluations of the performance of each of the filter used when applicable.

² <https://github.com/twintproject/twint>

3.1. First step: retrieve tweets related to the lexical field of inflation

This selection process relies on a dictionary-based filter build on six types of lexical fields related to prices problematics. If a tweet contains one of the keywords belonging to one of these lexical fields, it is selected. Otherwise, it is removed from the database.

As collected tweets are in French, most of the keywords are in French. However, since the language specified by Tweeter is automatically detected, some tweets might be in English. Moreover, some users may also use English expressions within tweets mostly written in French. For both reasons, a small set of keywords in English has also been defined. Table 1 displays the six lexical fields defined. It can be noted that the filter has been intentionally set broad to avoid missing any relevant tweet.

Lexical Field	Keywords originally used (French)	Keywords translated in English
<i>Lexical field of inflation with economical terms</i>	Inflation, déflation, stagflation, désinflation, inflationniste, déflationniste, antiinflationniste, antidéflationniste, IPC, IPCH	Inflation, deflation, stagflation, disinflation, inflationary, deflationary, anti-inflationary, anti-deflationary, IPC, IPCH
<i>Lexical Field of being expensive</i>	Onéreux, cher, prohibitif, couteux, élevé, exorbitant, inabordable, conséquent, inaccessible, excessif, anormal, dispendieux, arnaque, arnaquer, ruineux, faramineux, hors de portée, rondette, inconcevable, rédhibitoire	Expensive, expensive, prohibitive, costly, high, exorbitant, unaffordable, consequential, inaccessible, excessive, abnormal, expensive, rip-off, rip-off, ruinous, outrageous, out of reach, roundabout, inconceivable, prohibitive
<i>Lexical field of being cheap</i>	Faible, modique, avantageux, brader, imbattable, dérisoire, alléchant, réduit, occase, occasion, défiant toute concurrence, aubaine, modeste, clopinettes, bon prix, attrayant, clopinette, abordable, raisonnable, compétitif, accessible, acceptable, normaux, moyen, équitable, intéressant, convenable, négligeable.	Low, modest, advantageous, discounted, unbeatable, derisory, attractive, bargain, bargain price, attractive, bargain, affordable, reasonable, competitive, accessible, acceptable, normal, fair, interesting, suitable, negligible
<i>Lexical field of prices and costs</i>	Prix, tarif, montant, coût, loyer, vente, achat, location, frais, abonnement, facture, coûter, facturer, payer, tarifer, vendre, devis, paiement, rabais, tarifaire, croissance, promotion, remise, ristourne	Price, tariff, amount, cost, rent, sale, purchase, lease, fee, subscription, bill, cost, charge, pay, rate, sell, quote, payment, discount, tariff, growth, promotion, rebate, rebate
<i>Lexical field of statistical institutions related to the inflation's measure</i>	BCE, banque centrale, banque central, Banque de France, INSEE, FED, taux directeur, taux intérêt	ECB, central bank, central bank, Banque de France, INSEE, FED, key rate, interest rate

<i>Additional keywords in English</i>	Price, prices, cost, costs, rent, rents, bill, bills	
---------------------------------------	--	--

Table 1: List and content of lexical fields used to detect tweets related to prices matters

After this filtering procedure is applied, only 5% of the tweets remains, that is precisely 504,664 tweets. This lazy procedure needs not to be completed by a more demanding filtering, which will allow to remove the residual noise. Given the relatively small amount of remaining tweet, it is now possible to train a use a tailored supervised machine learning algorithm.

3.2. Second step: detect tweets related to prices

A supervised machine learning model is a model that provided numerical or categorical features of a tweet, will output a predicted label. It is called supervised because in the first place, it needs to be trained with a set of correct examples, *i.e.* of features associated with the correct label. In our case, this label is a binary value, which is set to 1 if the tweet is related to prices and 0 otherwise. In what follows, we explain the various components of our model: how the features of each tweet are constructed, what type of model is used and how it works, its performance, as well as how we manually labeled the training sample to train the model.

3.2.1. Explanatory variables: 200 word2vec variables

The goal of word embedding models, also called language models, is to create numerical vectors from textual data. Multiple word embedding models are available, and this study make use of the word2vec model (Mikolov *et al.*, 2013b), a probabilistic representation of words that take advantage of neural networks. In this model, the word is represented as a vector in the Euclidean space. Two words that appear frequently next to each other in the text will be close to each other in the Euclidean space. This proximity between the vectors can be measured using cosine similarity. In Appendix 7.1.3, we illustrate for instance what are the closest words from "price", "inflation" and "deflation" according to a word2vec model. Word2vec is a very popular and easy to use language model thanks to numerous implementations and an easy training that required no human labeling.

Many works have trained and used word2vec models, and consequently many pre-trained models are available online. It is relevant to use them when the data at hand is not sufficient to train correctly one's own word2vec algorithm. However, it can be interesting to train a word2vec model one's specific data when the database is large enough as in our case. In fact, most of the time, pre-trained models are trained on a huge amount of generic text data, such as Wikipedia articles or crawled websites. If the data at hand is specific to a topic or to a writing style, using such pre-trained models as is will result in poor performances. In this situation, training a new model better tailored to the data is often much better as it enables to capture the semantic specificities of the topic and style. In our study, the data at hand is specific because Twitter data is composed of short sentences about economics, hence with a specific style and vocabulary.

Therefore, we trained our own word2vec model on tweets. To make sure the model is trained on enough data, we use it on the full collected dataset of more than 10 million tweets.

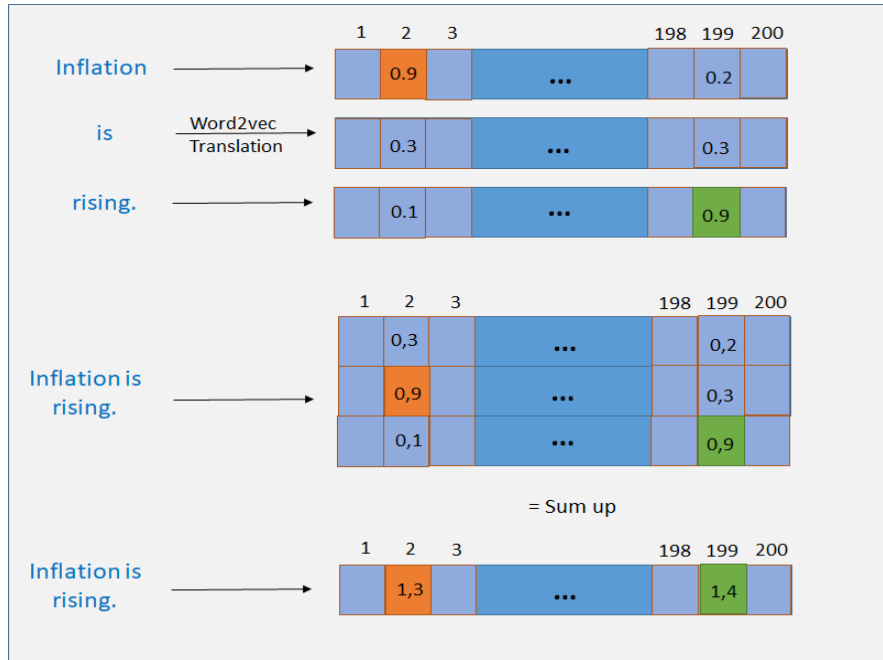


Figure 1 - Reading note: the diagram illustrates how the tweet "Inflation is rising" can be transformed into a vector with 200 coordinates. The three words "inflation", "is" and "rising" are transformed into a 200-coordinate vector. These coordinates characterize the word in question. For example, high values on coordinate 2 may indicate the presence of a price lexical field (orange color), and high values on coordinate 199 may indicate the presence of the increase lexical field (green color). The tweet can be represented as a matrix with 3 rows and 200 columns. To simplify this representation, we average over the rows: the tweet is simply characterized by 200 coordinates. Repeating this process for all tweets, each of them is characterized by 200 word2vec variables, representing as many lexical field signals more or less relevant for our analysis. (Values are given for explanatory purposes only).

In a word2vec representation, each word in the tweets is represented as a vector with K coordinates, as illustrated in Figure 1. The number of coordinates is a parameter of the model that needs to be chosen. This parameter is fixed by choosing the number of neurons that compose the hidden layer of the model. In our study, this number K was set to 200, as it is the case in most word2vec applications. Since the explanatory variable is to be computed at the tweet level, which are composed on several words, an aggregation method of each word's embedding must be applied. We chose to compute the average of the word2vec representations of each words in the tweet, along each dimension, resulting in a new 200 dimensions tweet embedding.

3.2.2. Explanatory variables: 38 additional dictionary-based variables

It is unclear, however, which of those 200 dimensions will (or will not) be related to the lexical field of price evolution. In order to add interpretability in the features used to characterize a tweet, we add 38 additional variables to each feature vector, which indicate the presence of 38 relevant lexical fields. For each tweet t , each indicator variable $X_{t,j}$ is set to 1 if the tweet contains a keyword belonging to the lexical field j , and 0 otherwise.

These lexical fields were built to detect a more precise notion among the topic of prices present in a tweet. They indicate the presence of characteristics in the tweet that can be grouped into five categories: the presence of words related to price topics or statistical institutions; the presence of words related to directional evolutions of price; the presence of words related to the lexical field of "cheap" or "expensive"; the presence of degree adverbs and adjective or negation terms; and the presence of words to be excluded (false friends of keywords that belong to the lexical field of prices).

The majority of the variables rely on French keywords whereas seven variables rely on English ones. In further developments, we could create more variables based on the English vocabulary. The variables also distinguish between verb and noun in order to integrate grammatical features into the model. In Appendix 7.5, we detail further the keywords used for each variable.

3.2.3. The labelling process

To train the model, we manually labelled a random sample of 800 tweets. As a recall, label 1 was assigned if the tweet is related to prices and 0 otherwise. This quite straightforward task required no specific knowledge, but it was however realized by trained economists to ensure consistency. Table 3 shows a typical example of the obtained labeled dataset. In a second step, which will be explained further, we tagged each tweet identified as "about price" according to one of the five following categories depending on its precise content: "deflation", "inflation", "disinflation", "price stability" or "unspecified". The "unspecified" label is important because most of the tweets mentioning prices actually do not mention any price evolution or perception of such evolution. For more information about the labelling process, please refer to Appendix 7.3.

Original text (French)	Translated text (for paper purpose)	Label
L'OPEP veut voir les prix du pétrole revenir à un niveau « raisonnable »	OPEC wants to see oil prices return to a "reasonable" level	1
Prix d'excellence Alassane Ouattara 2018	Alassane Ouattara Excellence Award 2018	0

Table 2: Typical result of the labelling process (without categories)

3.2.4. The model: random forest

Using the resulting 238 obtained features computed for each tweet, we train a random forest model to detect tweets whether a tweet is related to prices matter. This type of model architecture is well tailored for classification and is particularly relevant with high dimensional data. Its specificity lies in combining a multitude of decision trees, and outputs the class obtained from a majority vote of all decision trees. This methodology has also the advantage to produce little overfitting. Appendix 7.2 offers additional information about random forests and its functioning.

To optimize a random forest, two key parameters need to be tuned: the number of decision trees generated to include in the random forest, also called *ntree*, and the number of explanatory variables that will be considered at each tree split, also called *mtry* (see Kern, 2019). In general, the number of trees is set to 500 and the number of variable to consider at each split at the square root of the number of explanatory variables used in the random forest. However, these values are only starting points and they need to be calibrated in function of the data at hand and the problem we need to model. A large set of values for these parameters – between 50 and 2500 for *ntree* and between 0 and 30 for *mtry* – have been tested using cross-validation.

Figure 2 reports the result of this hyper-parameters optimization. After a certain point, increasing *ntree* or *mtry* does not improve significantly the performance, according to both accuracy and kappa metrics. In the case of *ntree*, a value greater or equal to 500 seems to stabilize the performance, but increasing this value also increase the computation time of the algorithm. After analyzing these results, we decide to set *mtry* to 30 and *ntree* to 500. Appendix 7.6.2 offers more information about these evaluation metrics.

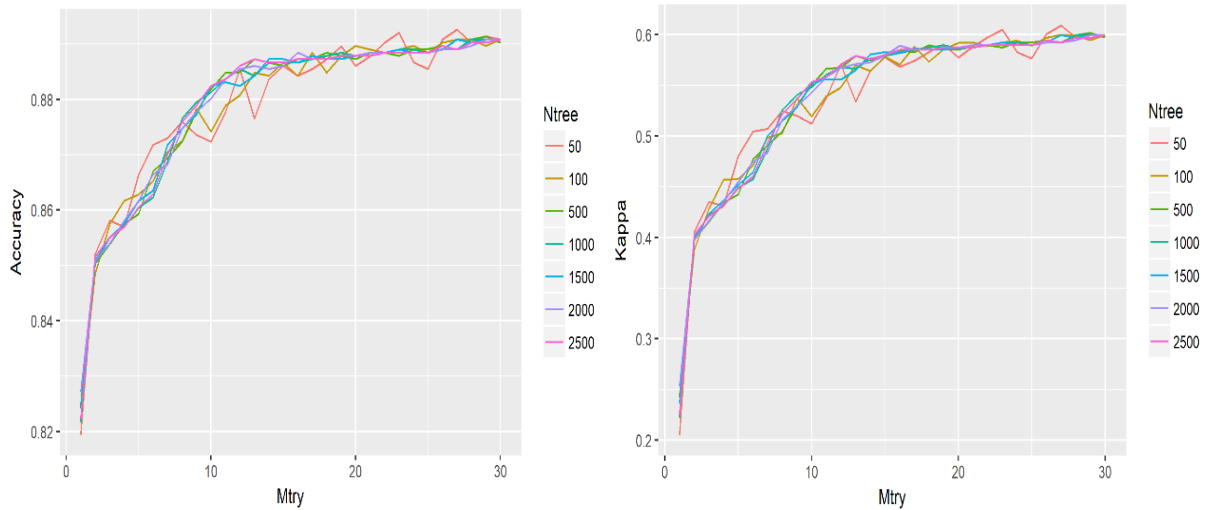


Figure 2: Optimizing the random forest with ntree and mtry

3.2.5. Setting the probability threshold

The raw output of such random forest is a predicted probability that a tweet is related to prices. To classify a tweet based on this probability, it is necessary to set a threshold that will act as a frontier: tweets with a probability above this threshold will be considered as talking about prices, otherwise they are considered as off-topic. To determine the optimal threshold, we compute the false positive rate and the false negative rate for different threshold value. A higher threshold value has two opposite effects on these rates (Figure 3). First, the false positive rate decreases, since the probability to belong to the "on topic" class predicted by the model has to be higher and higher so that only the "most certain" ones remain; and second, the false negative rate increases, since the model is more selective and misses more tweets related to prices matters.

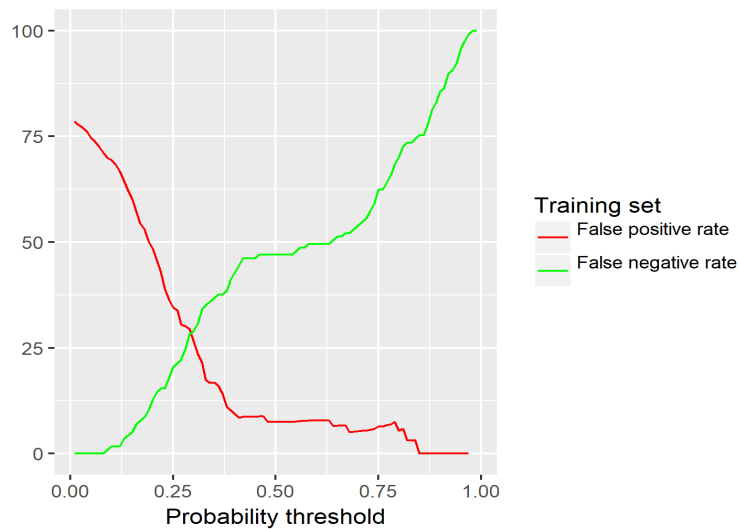


Figure 3: Setting the probability threshold

We chose to select the threshold in such a way as to obtain as many false positives as false negatives. For our study, these two types of error have a priori the same importance and are both to be minimized. We do not want to decrease one at the risk of increasing the other. The probability threshold is therefore set to 0.3, since is exactly the point where the false positive rate equals to the false negative

rate in our training data as seen in Figure 3. In other words, it means that the model classifies all the tweets with a probability greater than 30% in the category "related to price matters", and to the "off-topic" category otherwise.

3.2.6. Variable importance

One key advantage of random forests is that it makes it possible to identify which features contribute the most to the predictions by computing the variable importance as the average decrease in Gini metric. Each node that composes a decision tree in the random forest is a condition based on a single predictor that split the data in two datasets (for instance, split observations into those having feature $k \geq 0.5$, and those having $k < 0.5$). The "Gini impurity" is the most often used metric to get this optimal (local) condition. It is possible to calculate afterwards by how much each variable decreases the average "impurity" of the tree in total. In the case of a random forest, the importance of each predictor is obtained by averaging the impurity in each tree of the forest.

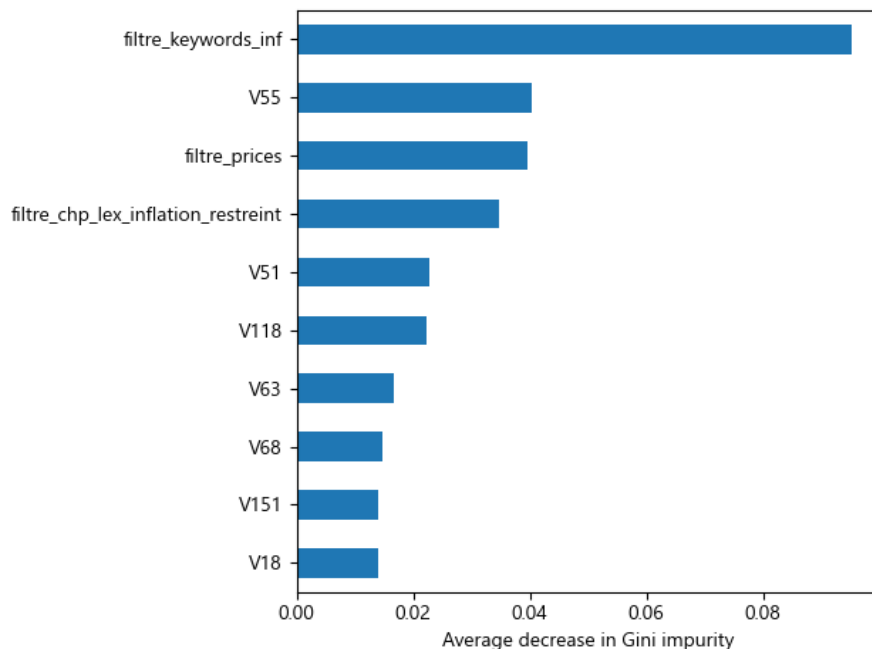


Figure 4: Gini feature importance

Figure 4 displays the obtained feature importance analysis for our model. Both word2vec and dictionary-based features seem to contribute significantly to predictions. Using these two kind of variables was justified as it provides relevant information. However, these results should be interpreted cautiously: other variables may also be important and still not appear on the figure because of multicollinearity between variables.

3.2.7. Performance evaluation

The performance of the random forest can be assessed using the area under the ROC curve (AUC) metric on train (560 tweets, 70% of the labeled data) and test set (240 tweets, 30% of the data). An AUC metric much higher on the test set than on the train test could be a sign of overfitting. As displayed on Figure 5, our model shows little overfitting as the ROC curve for train and test sets are similar. Overall, the model performs very well: the values for the AUC metrics are very high for the train and test set.

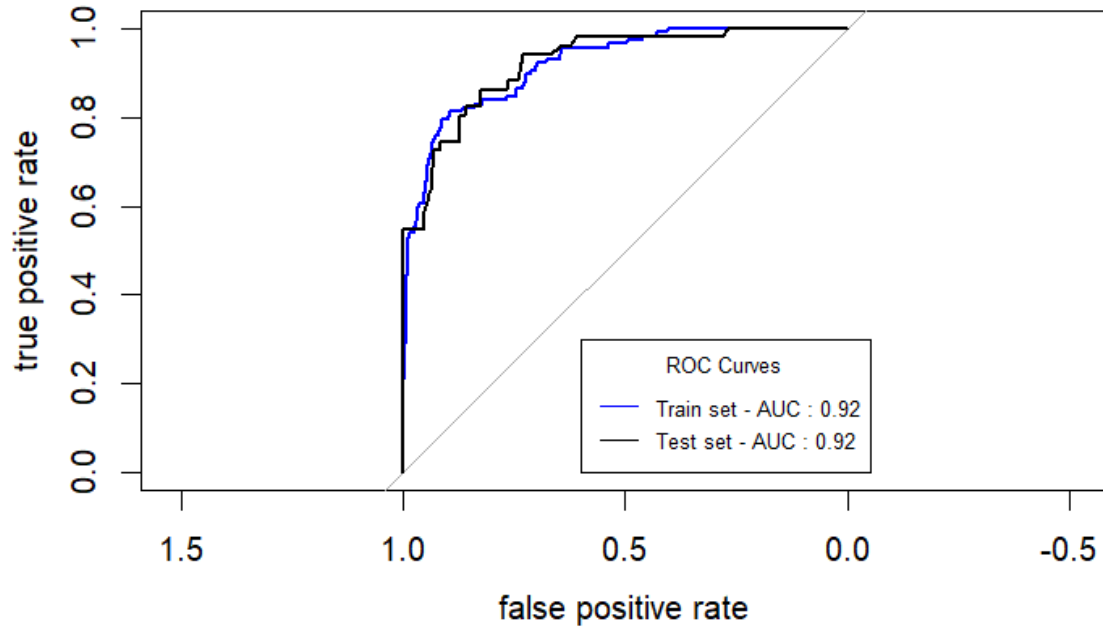


Figure 5: Comparing ROC curves between train and test sets

Other evaluation metrics, displayed in Table 5, also demonstrate performance of the model. We computed the metric both on the training set and on the test set, using standard evaluation metrics. Please refer to Appendix 7.6.2 for more information about their definitions. Obtained values are close to each other, which indicates little overfitting. To assess the quality of the model, it is necessary to consider and compare multiple metrics because the response variable is significantly imbalanced. This imbalanced data increases artificially the AUC or the accuracy. We rather use precision, recall and F-score, which are better tailored to handle such situation. These two indicators are indeed slightly lower but still high enough to warrant a good performance of the model.

Metrics	Training sample	Test sample
Accuracy	0.912	0.879
F1-score	0.819	0.739
Precision	0.817	0.732
Recall	0.822	0.746

Table 3: Evaluation metrics of the model

3.3. Third step: identify the direction of price mentioned

In previous steps, we identified tweets related to prices in general using keywords and machine learning methods. In this last step, our goal is to go one-step further and categorize what opinion about price evolution each tweet conveys. We define four categories (or topics) of interest in which each tweet could fall: inflation, deflation, disinflation and prices stability. However, a preliminary analysis shows that even among tweets related to prices, almost half of them do not match any type. We therefore add an additional "Other" category, in which fall all tweets not mentioning any of the four topics of interest. Table 3 displays an example of tweet (translated in English) for each type of content.

Tweet text (translated for paper purpose)	Category
Fukushima: electricity prices increase lead to 10 times more deaths than the accident itself.	Inflation
Another one bites the dust... #retailapocalypse #diesel #realestate #lifestyle #disruption #deflation	Deflation
Euro Area inflation expectations keep dropping	Disinflation
So one of the euro's concrete objectives, to control inflation, has been achieved.	Stability
Are you broke? Your banker is rubbing his hands. Cost per client: 58.91 euros.	Other (out of topic)

Table 4: Illustration of the classification of the tweets regarding prices evolution

As mentioned earlier, we annotated a sample of 800 tweets, both with a binary label to identify whether they are related to prices, but also according to one of our five topics if they were. However, our initial ambition to train another supervised machine learning model to automatically recognize the topic mentioned cannot be achieved with our annotated database. In fact, with only 168 tweets being tagged as related to prices, and then about a few dozen of them falling into the "Other" category, there remains less than a dozen example for each category of interest. Labelling more tweets to get significantly more training data would be necessary to train such a classifier. Further development will undertake this task.

For this reason, we chose to use a keywords-based method to identify the direction of prices mentioned in the tweets. Appendix 7.4.2 presents the combination of lexical fields used, which focused on the hard task to disentangle the thin differences of some of our four topics of interest. We make the hypothesis that, at this point, using keywords related to the increase, decrease, and stability's lexical fields could be enough to identify whether a tweet is talking about inflation, deflation, disinflation or prices stability, as most of the noise has already been ruled out. Further development will focus on improving the reliability of this last task.

3.4. Last step: excluding tweets mentioning foreign prices

Despite our focus on French perception of inflation and thus tweets about French prices, a few tweets explicitly concerning prices of foreign countries remain in our database. In the labelled dataset, only 3% (24 tweets) of the tweets were concerned, yet we decided to exclude such tweets from our general database using simple rules. First, we remove tweets that mention a country name, in French or English, other than France. Second, we remove tweets that mention a nationality name, also in French or English, other than "French". Note that we excluded tweets that mention global or European prices, as we attempt to keep a certain purity in the index despite the probable correlation between European and French prices evolution.

3.5. Review of the quality of this four-step methodology

It is possible to estimate the overall quality and the performance of our methodology. The aim is to estimate the quality of the detection of the tweets falling in our four categories of interest, obtained by a two-step classification with machine learning followed by a keywords-based method. We compute

the metrics using our database of 800 labeled tweets despite the little counts in each category. Table 5 displays those metrics.

Category	Precision	Recall	F-score
<i>Inflation</i>	0.64	0.56	0.60
<i>Deflation</i>	0.78	0.97	0.86
<i>Disinflation</i>	0.00	0.00	0.00
<i>Stabilization</i>	0.56	0.71	0.63
<i>Other</i>	0.78	0.86	0.82

Table 5: Overall model evaluation

Results can be read as followed, in the case of the inflation topic. Among tweets returned as “about inflation”, 64% are indeed about inflation (recall) and 36% are false positive. On the other hand, when considering all tweets truly about inflation, 56% are identified as such by our methodology, and hence 44% mentioned are missed and not classified as “about inflation” (recall). All results read the same for each category, with the notable case of disinflation where all scores as set to zero. This can be explained by the fact that only 16 examples are labeled as such in our training dataset, and therefore it is hard to obtain statistically significant results.

Overall, metrics show that our methodology has satisfying success when it comes to identify tweets according to their topical content; yet the score greatly depends on the category. The methodology for instance succeeds in detecting tweets related to deflation and those falling in the out-of-topic category, but encounters more difficulties to detect tweets about inflation, disinflation or prices stability. This highlights the limit of using just keywords in the last step of our methodology. Once again, this highlights the necessity to improve this last step in future works.

4. Results

4.1. Twitter indicator of perceived inflation

The methodology detailed in the previous section enables us to identify tweets mentioning the problematic of price, and then to detect tweets related to inflation, deflation, and other minor topics. The indicator we propose builds on this methodology with a simple computation. We simply count, for each given period (day, week, or month) the number of tweets mentioning inflation, and the number of tweets mentioning deflation. Figure 6 displays both those counts at a monthly scale. Inspired by the work of Angelico *et al.* (2021), the indicator is the number of tweets about inflation minus the number of tweets about deflation. Note that we do not take into account tweets mentioning disinflation or price stability, mainly because the quality of the detection for those categories lack precision. We also compute a smoothed version of the indicator, using a backward-looking exponential weighted moving average with parameter α set to 0.35.

It can be noticed that from the huge volume of collected data, with more than 10 million tweets, only a minority pass all filters and end up will contributing to the construction of the Twitter indicator. Each month, about 13,000 tweets are published on average by Banque de France retweeters, but only 58 relate to inflation and 27 to deflation.

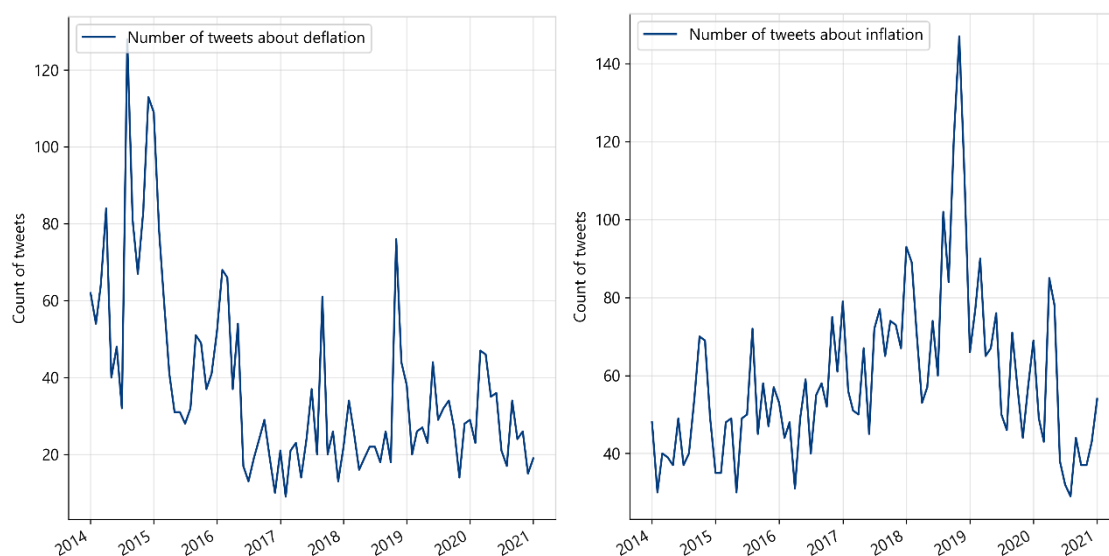


Figure 6: Comparing the number of tweets mentioning deflation or inflation over time

The Twitter indicator and its smoothed transformation appear on Figure 7. The indicator is negative just at the beginning of the period, due to the spike of tweets talking about deflation at the beginning of the year of 2015 as seen in Figure 6(a). To assess further the quality of the index, it is also necessary to compare it with other measures of perceived inflation, as well as the true inflation rate.

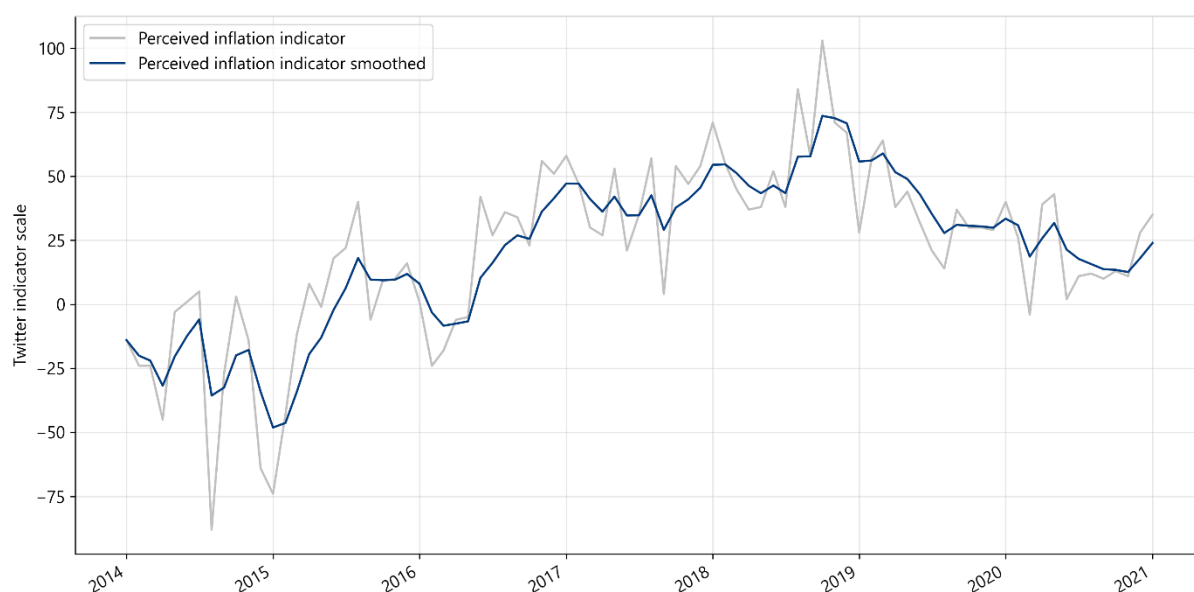


Figure 7: Twitter indicator

4.2. Twitter indicator consistency with households surveys

The French National Institute for Statistics (INSEE) produces its household survey each month, whereby it questions a sample of French households about their opinion on their economic environment.³ Among other subjects, respondents are asked about their perceptions of inflation and their anticipations on its evolution. Respondents are asked to choose between “increase” and “decrease” on the two following questions: “how do you think prices have evolved over the past 12 months?” and:

³ Data are available at: <https://www.insee.fr/fr/statistiques/series/102414547?INDICATEUR=2874666%2B2874667>

“how do you think prices will evolve over the next 12 months?”. INSEE constructs two indicators from those answers: a past evolution indicator, which is the percentage of households that think prices have increased over the past 12 months, and a future evolution indicator, as the percentage of households that think prices will increase in the next 12 months. Both can be used as reference indicators as the overall French population’s perception of inflation.

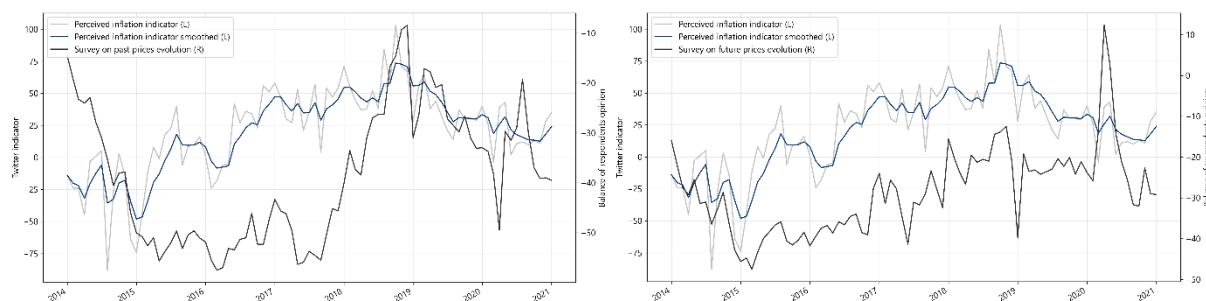


Figure 8: Households’ perceived evolution of past prices (left) and future prices (right) compared to Twitter indicator

Our findings show that the Twitter indicator is correlated with both INSEE indicators (past and future), using both the Pearson and Spearman correlation coefficients, as shown in Table 4 (see Appendix 7.6.1 for more details about both metrics). The correlation is notably stronger when the Twitter indicator has been smoothed. In Figure 8, we display both INSEE indicators along with our Twitter index. We see that in both cases, curves seem to follow the same trends, but also peak at the same time. Our index therefore seems consistent with those reference indicators. An interrogation arises on whether the Twitter indicator measures anticipation of future inflation, or the perception of past inflation. Since correlation coefficients are always higher with the past evolution INSEE indicator than with the future evolution one, we can hypothesize that it is rather related to anticipation of future inflation.

Correlation between...	Pearson	Spearman
Past evolution perception and Twitter indicator	0.213	0.219
Past evolution perception and smoothed Twitter indicator	0.305	0.329
Future evolution perception and Twitter indicator	0.460	0.487
Future evolution perception and smoothed Twitter indicator	0.505	0.558

Table 7: Correlation summary Table of the INSEE statistics and the Twitter indicator

4.3. Twitter indicator consistency with the inflation rate

INSEE also publishes the Consumer Price Index (CPI), which is the base for the calculation of the inflation rate as the evolution of the variation over month of the CPI.⁴ Table 8 shows that the Twitter indicator is strongly correlated to the inflation rate, and Figure 9 displays the two series together. Pearson and Spearman correlation metrics are indeed high and always between 0.6 and 0.8. The correlation is once again stronger when the Twitter indicator is smoothed.

Correlation between...	Pearson	Spearman
Inflation rate and Twitter indicator	0.642	0.664

⁴ Data are available at: <https://www.insee.fr/fr/statistiques/serie/001763852>

Table 8: Correlation summary Table of the inflation rate and the Twitter indicator

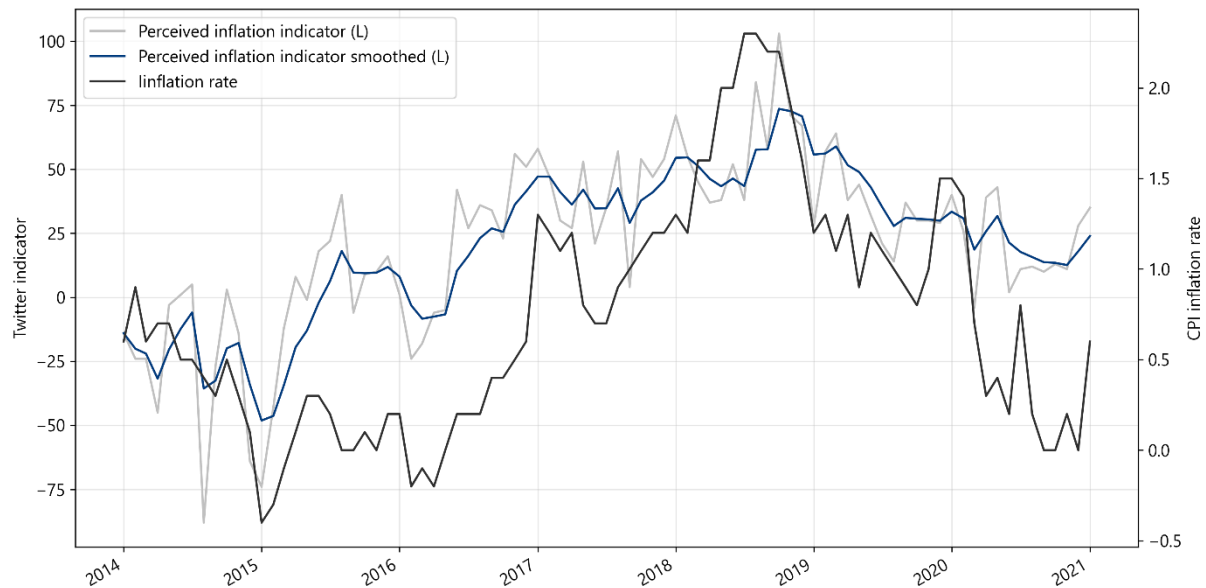


Figure 9: Twitter indicator and true inflation measured with CPI change

One interesting question is also to investigate whether our indicator can be anticipatory of the future inflation rate, or rather that it reflects past inflation rate. To do so, we analyze the time-lagged correlations between the inflation rate and the Twitter indicator. Figure 10 displays the Pearson and Spearman correlations' variations when the Twitter indicator is compared to the inflation rate in a range from 12 months earlier to 12 months later. Here is how to read the leftmost point: the Pearson correlation between the Twitter indicator at month m and the inflation rate at month $m-12$ (a year before) is about 0.195. The rightmost point reads similarly: the Spearman correlation between the Twitter indicator at month m and the inflation rate at month $m+12$ (a year later) is about 0.580.

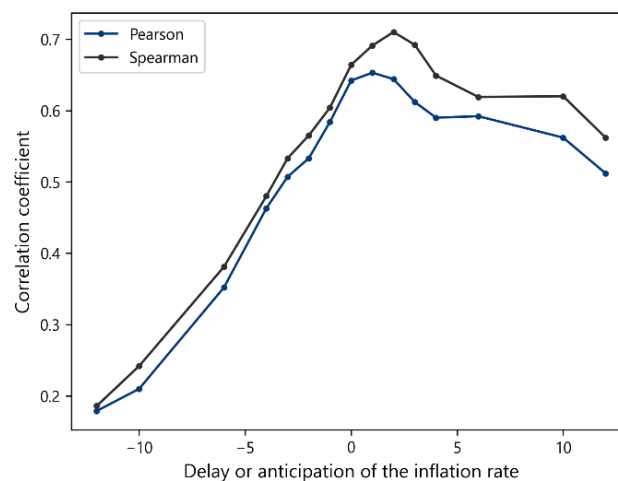


Figure 10: Time-lagged correlations between the smoothed Twitter indicator and the inflation rate

A clear conclusion appears from this plot: the Twitter indicator is much more correlated to the future inflation rate than to the past inflation rate. Interestingly, inflation rate 12 months in the past has a low correlation coefficient of 0.195, but inflation rate 12 months in the future keeps a high correlation coefficient, between 0.5 and 0.6. More interestingly, the indicator is more highly correlated to the

inflation rates of month $m+1$ and month $m+2$ than it is with the one of month m . This clearly indicates that our Twitter indicator is a forward-looking measure, which provides indications about *perceptions of future of inflation* (i.e. anticipations) rather than *perceptions of past inflation*.

4.4. Twitter indicator and Covid-19

A robustness check of our indicator can be conducted by studying its behavior during the beginning of the Covid-19 economic crisis. Interestingly, the indicator is quite stable during this period. In fact, 30% more tweets have been posted during the French lockdown period (March to May 2020) than before March 2020, and a significant amount of them concerns the pandemic: 16% of the tweets posted between March and May 2020 are related to the Covid-19 disease⁵. It also had an effect on the number of tweets concerning inflation or deflation matters, which increased by 28% during the first French lockdown compared to the previous period.

Period	Number of tweets (monthly)	Number of tweets about Covid-19	Number of tweets about deflation or inflation
<i>Before the lockdown of March 2020</i>	13 450	80	92
<i>During the lockdown (March to May 2020)</i>	17 296	2 781	125

Table 6: Comparing tweets volume before and during the French lockdown of March 2020

Interestingly, it seems from Figure 6 that tweets related to inflation and those related to deflation counterbalance each other, resulting in a stability of the Twitter indicator during that period (see Figure 6). This behavior is coherent with the behavior of the true inflation rate, which has increased in some sectors (food prices for instance), and decreased in other sectors (e.g. energy), resulting in a global inflation rate remained globally stable during this period (see INSEE, 2020).

4.5. How do people talk about prices?

Word clouds enable us to visualize words that appear most often in the tweets and thus contribute the most to the Twitter indicator. Then, particular topics can be highlighted. Figure 11 displays the word cloud on the subset of the tweets related to either deflation or inflation topics.

Naturally, words like “prices”, “price”, “inflation”, “costs”, “bill”, or “prix” (French word for “price” and “prices”) are the most represented. They belong to the set of keywords used in the first place to filter tweets. What has more value is to analyze words that do not belong to the lexical field of price evolution. Some sectors come up more often than others do. In particular, the oil (“oil” or the French word “pétrole”), the housing (“housing” or the French word “logement”, “rents” or the French word “loyer”, “immobilier” which means “real estate” in French), and the energy (“energy”) sectors. It could be interesting to isolate these tweets to build an indicator measuring prices on these particular sectors. The area of Paris is also more mentioned than others. Some major events that are likely to impact prices such as “Brexit” or “Covid-19” pandemic are also present in the word cloud.

⁵ Keywords used are to detect them are “covid”, “coronavirus”, “pandemic”, and “epidemic”, both in English and French.



Figure 11: Word cloud of the tweets behind the Twitter indicator

5. Conclusion and future works

Finding the tools to classify and evaluate the methodology was one of the main challenges of the study. Combining machine learning methods and keywords filtering was a good way to fulfill those tasks and provided good results. The performance metrics of the classifier are high and the resulting Twitter indicator seems to provide useful information. The indicator is indeed consistent with the monthly household surveys on inflation expectations and is highly correlated with the inflation rate. Overall, our study demonstrates that it is both possible and relevant to use Twitter data to measure inflation perceptions. The results provide reassurance about the bias in the data, restricted to the Banque de France retweeters. Measuring the perception of specific users proved to be an appropriate approach. In addition to enable profile characterization, restricting the study to expert profiles gives relevant result because their perception seems to be more accurate of the reality by being based on scientific grounds.

Of course, there are still open questions that remain and motivate further developments. First and above all, the data is restricted to a small panel of users who post rarely about prices matters despite their expert profile. It would be most useful to expand the initial data at hand by extending the number of expert profiles analyzed. Three main approaches for instance be explored: following a social community approach by integrating the contacts of users used in this version of the indicator, including the retweeters of the European Central Bank, or even use the list of predetermined economists who have a Twitter account. Of course, the most complete approach would be to include all tweets mentioning prices in French, irrespective of the users, as to obtain a fuller picture. This raises of course the question of the volume of the data that would be collected, and would require computing tools accordingly.

The quality of the various classification and filtering steps could be improved by extending the labelled database, which would exempt us from using keywords to categorize the content of the tweets regarding prices evolution. This would make it possible to train a supervised machine learning model to classify the content directly and not just detecting tweets related to prices matters as in this study.

Finally, no normalization is actually involved in the Twitter indicator. This was a choice driven by the idea that an increase of tweets about inflation was still an interesting signal to measure even though the total number of posts also increased. However, testing several types of normalizations to capture additional information would be relevant. For instance, one could think of normalizing by the sum of number of tweets talking about inflation, deflation or off-topic, as this would have the merit of taking into account "off-topic" (or rather, "neutral") tweets. Another idea could be to normalize by the total number of number of tweets that mention economic topics, not just inflation. This normalization would make it possible to measure the relative importance of inflation within economic topics, but would require properly defining what an economic topic is.

Finally, a crucial flaw in this work is that we do not yet make the distinction between tweets expressing a personal opinion about inflation and those that merely reflect official announcements of statistical institutions about inflation. This "sounding board" effect is interesting to study, particularly with regard to the impact of institutional communication, but it may also introduce biases to the measurement. It should therefore be more clearly investigated and the two effects should be separated.

6. References

- Altig D., Baker S., Barrero J. M. , Bloom N., Bunn P., Chen S., Davis S. J., Leather J., Meyer B., Mihaylov E., Mizen P., Parker N., Renault T., Smietanka P. and Thwaites G. (2020). "[Economic Uncertainty Before and During the COVID-19 Pandemic](#)". *Journal of Public Economics* 191.
- Angelico C., Marcucci J., Miccoli M., and Quarta F. (2021). "[Can we Measure Inflation Expectations Using Twitter?](#)". *Banca d'Italia Temi di discussione* n°1318.
- Baker S. R., Bloom N. and Davis S. J. (2016). "[Measuring economic policy uncertainty](#)". *The Quarterly Journal of Economics*, 131(4).
- Bec F. and Mogliani M. (2013). "[Nowcasting French GDP in Real-Time from Survey Opinions: Information or Forecast Combinations?](#)". *Banque de France Working Paper Series* n°436.
- Bertoli C., Combes S., Renault T. (2017). "[Comment prévoir l'emploi en lisant le journal](#)". *Note de conjoncture de l'INSEE* of 03/2017.
- Kern C. (2019). "[Tree-based Machine Learning Methods for Survey Research](#)". *Survey Research Methods* 13(1).
- Kintzler E. (2018). "La Banque de France sur Twitter : impact des publications et réseaux d'influence". *Banque de France Internal Research Document* n°18-025.
- Mikolov T., Chen K., Corrado G., and Dean J. (2013a). "[Efficient Estimation of Word Representations in Vector Space](#)". *Proceedings of Workshop at ICLR 2013*.
- Mikolov T., Sutskever I., Chen K., Corrado G., and Dean J. (2013b). "[Distributed Representations of Words and Phrases and their Compositionality](#)". *Proceedings of NIPS 2013*.
- Thorsrud L. A. (2016). "[Nowcasting using news topics, Big Data versus big bank](#)", *Norges Bank Working Paper* n° 20/2016.

7. Appendix

7.1. Word2vec model

7.1.1. How it works

The Word2vec model is a word embedding method based on a probabilistic representation of words and on neural networks. It has made it possible to rethink the concept of word embeddings by representing words in a vector space where words used in similar contexts are represented close to each other. The Word2vec representation of a word depends indeed on its "context", it means it depends on the words surrounding the term considered in the sentences of interest. The similarity between two word-vectors can be measured with the cosine similarity metric, a little further.

The word2vec model relies on a two-layer neural network. Two types of neural architectures can be used. In the Continuous Bag of Words (CBOW) architecture, the neural network tries to predict a word according to its context. In the Skip-Gram architecture, the neural network tries to predict the context according to the given word. In both cases, the neural network takes unstructured text as input, and modifies its neural weights using unsupervised learning to reduce the prediction error of the algorithm. It is possible to fix the number of word-vector coordinates obtained by the model by choosing the number of neurons number of the hidden layer.

The word2vec model has multiple advantages. For its training, word2vec only needs raw text data that do not require to be labelled. Therefore, a large corpus of unstructured set is enough to estimate a model with good performance. Finally, the algorithm is efficient and can be run on a huge volume of data in a minimum of time. This is mainly due to its simple neural network structure.

7.1.2. Optimization of hyper parameters

Many hyper-parameters can be tuned to improve the model performance. We highlight three of them. The *dimension of the vector space* it is the number of numerical predictors used to describe the words (between 100 and 1000 in general), in other words the number of coordinates characterizing the vector representation of a word. The *architecture of the neural network* is a second parameter. It must be chosen between Continuous Bag of Words (CBOW) and Skip-Gram. Finally, the *size of the context* is a last crucial parameter. It refers to the number of terms surrounding the word in the sentence. According to the creators of word2vec, it is recommended to use contexts of size 10 with the Skip-Gram architecture and 5 with the CBOW architecture (see Mikolov *et al.*, 2013a).

7.1.3. Application in this study

As explained on the paper, many pre-trained word2vec models are freely available online. They are even more useful when the volume of data at hand is not sufficient to train correctly the Word2vec model. Most of the time, pre-trained models rely on a huge volume of generic data. Therefore, the resulting word-vector representations are also generic. In the case of specific data, training the model on this data can be judicious as it allows capturing semantic relations specific to the field of study. With its short sentences, Twitter data is atypical and fits in this kind of use case.

Training a Word2vec model on one's own data can be done under Python with the *Gensim* package or under R with the *wordVectors* package. In our study, we trained a word2vec model on tweets

using Python and thus Gensim. To train it on enough data, all tweets collected were used for the training, before any filtering was applied. In the end, the Word2vec model has been trained on more than 200 million French and English words. The hyper-parameters chosen are listed in Table 7.

Parameter	Value
<i>Dimension of the vector space</i>	200
<i>Context size window</i>	5
<i>Number of times to process the entire corpus for training</i>	100
<i>Type of neural architecture</i>	CBOW
<i>Minimum times a word must appear in all tweets to be included in the training process</i>	5

Table 7: Hyper-parameters for the word2vec model of the study

After training, to check if the Word2vec representation is coherent, it is possible to look at the 10 words that are the closest to terms specific for our analysis. In Table 8, we do so for words “prix” (“prices” and “price” in English), “inflation”, and “deflation”. We provide words in French along with their English translation when needed.

Rank	“Prix”	“Inflation”	“Deflation”
1	croissance (<i>growth</i>)	fed	deflationniste (<i>deflationist</i>)
2	achat (<i>purchase</i>)	recession	inflation
3	tarif (<i>rate</i>)	infl	recession
4	cout (<i>cost</i>)	insee	fed
5	moyen (<i>means</i> , probably for <i>means of payment</i>)	hicp	easing (probably for <i>quantitative easing</i>)
6	loyer (<i>rent</i>)	eurozone	spiral
7	cher (<i>expensive</i>)	contraction	bulle (<i>bubble</i>)
8	frais (<i>fees</i>)	ipc (<i>cpi</i>)	qe (for <i>quantitative easing</i>)
9	occasion (<i>second-hand</i>)	evaporation	deflationnaire (<i>deflationary</i>)
10	vendre (<i>sell</i>)	draghi	eclatement (<i>bursting</i>)

Table 8: Closest words to relevant terms for the study

This example demonstrates that the word2vec representation has indeed captured the context of terms that usually go along with a word: the closest words belong indeed to the lexical field of inflation, prices and deflation context of words.

7.2. Random forest

The random forest model is appropriate for categorical (so-called “classification” problem) or numerical outcome variables (so-called “regression” problem). Input predictors (or features) can be both categorical and numerical variables. This methodological appendix concerns the case where the response variable is categorical (and even binary, as in this study).

Random forests are the result of combining a multitude of decision trees from the CART algorithm. A decision tree creates multiple partitions of data using a set of rules to predict the class of each observation. Decision trees are intuitive and interpretable algorithms but they also tend to overfit data. Random forests solve this issue. The bagging approach is used to generate the trees. Sub-samples are generated using a random sampling without replacement. Then a CART-type algorithm is applied on each sub-sample to build a decision tree. Not all variables are input of the algorithm but a random sample without replacement. These samplings enable to build independent decision trees being trained with different observations and different variables. That is why random forests tend to present less overfitting and are better generalized on new data.

7.2.1. Hyper-parameters optimization

To optimize random forest, two key parameters have to be considered: the number of decision trees (*ntree*) and the number of explanatory variables that each decision node will take as input (*mtry*). In general, *ntree* is set at 500 and *mtry* is set at the square root of the number of predictors used in the random forest. However, these values are only starting points and they need to be calibrated in function of the data at hand and the problem we need to model. In the case of *ntree*, increasing this value also increase the computation time of the algorithm. That is why we rather keep this value not too high.

7.2.2. Random forests output

A random forest produces for each observation a prediction. For a classification problem, each tree computed by the random forest gives a prediction - in our study, it would be "the tweet is related to prices matters" or "the tweet is not related to prices matters". The final prediction is computed as the most frequent prediction among all predictions produced by all trees.

A random forest can also produce the probability for an observation to belong to a class. This is what is used in our study. The computation of probabilities actually depends on the implementation of the random forest algorithm. In our study, we used the *scikit-learn* package on Python, which provides the proportion of decision trees classifying a tweet as related to prices matters.

7.2.3. Variables importance

Inspecting the importance of the variables is crucial to determine which predictors have the most contributed to the predictions. In this perspective, two measures are possible in the classification case. First, Mean Decrease Accuracy, which is constructed by swapping the values of a given predictor and look at the impact produced by calculating the decrease of accuracy. The more important predictors are, the more significantly the accuracy of the model decreases.

Second, Mean Decrease Gini. In a decision tree, each node that composes a decision tree is a condition based on a single predictor. To get an optimal (local) condition, the metric often used is the "Gini impurity". When training a tree, it is possible to calculate the impact of each variable on the average "impurity" of the tree. At the scale of a random forest, it is possible to calculate the importance of each predictor by averaging the "impurity" obtained for each tree of the forest.

Inspecting the importance of the variables is crucial to understand the relationships between the explanatory variables and the response variable. By being an interpretable model, the random forest provides a descriptive analysis of the data. However, a prerequisite to a good analysis is the absence of collinearity between variables. This issue is not a real concern in a predictive approach but it prevents

from analyzing causal links. Indeed, when two (or more) variables are strongly collinear, only one of the variables is more likely to capture all the importance. The importance of the other variables is somehow "hidden" by the one capturing all the importance. For this reason, we have not done a deep analysis of the variables importance in our study. The variables importance mostly gives information about the forest construction and does not really enable us to deduce any causal links. To do such causal analysis, a more in-depth study on collinearity links within the predictors would have been necessary. This could be a line of research for further developments

7.3. Description of the human labelling process

To train a machine learning model able to detect when tweets relate to prices matters, we randomly selected a sample of 800 tweets among the tweets containing at least one of the keywords related to the lexical field of prices. Then, we labelled this sample to train a supervised machine learning model to detect tweets related to prices. The manual annotation process consisted in creating four variables by reading the text of the tweet. Not all of them are used in the project, but may be useful in future developments.

Variable *"is_inf"*: asked the question: "is the tweet related to prices matters?". It is a binary variable, which is assigned to 1 if the answer is yes, 0 otherwise. This variable enables us to detect tweets related to our problematic. Despite having used a preliminary dictionary-based filter, only 21% of the tweets are found to concern prices in our training dataset.

Variable *"what_info"* answered to the following question, if a tweet was first labeled as being related to prices: "what is its content?". Possible answers were one of our five categories of interest described in the paper: "inflation", "disinflation", "deflation", "prices stability", or "other" (out of topic). The variable helps to describe the content of the tweet regarding the prices problematic.

Variable *"what_prices"* asked, "if the tweet is related to prices matters, what kind of prices are mentioned?". Two answers were anticipated either "global prices" or "prices concerning particular sectors". This variable allows determining if a significant part of the tweets concerns particular sectors. Almost half of the tweets classified as talking about prices matters concern indeed particular sectors.

Variable *"what_loc"* was finally concerned with the following question: "if the tweet is related to prices matters, what geographical localization is it referring to?". The goal was to filter tweets that explicitly mentioned something else than France, rather than having a precise location. Its possible values are "Global/French prices" or "prices concerning explicitly another country". Around 15% of the tweets related to prices matters concern explicitly prices from another country. These later tweets were removed from our database.

Finally, Figures 12 to 15 describe the counts of each modality for our four variables in our labelled database of 800 tweets.

7.4. Dictionary-based filters

In this study, we used several kind of filters relying on keywords. The tables below explain the lexical fields used and their corresponding keywords. The lexical fields and dictionary-based filters have been set with an expert perspective and by confronting multiple examples of tweets.

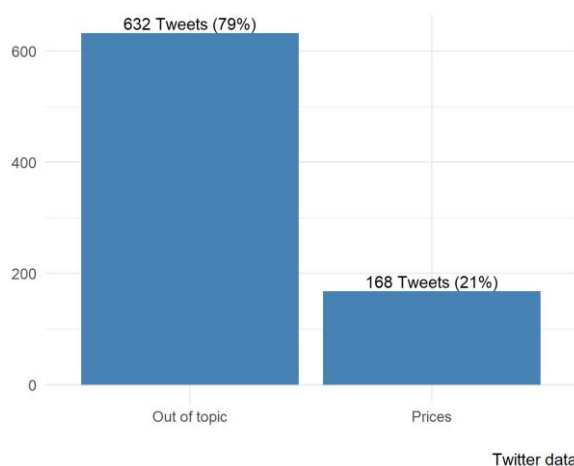


Figure 12: Description of variable *is_inf*

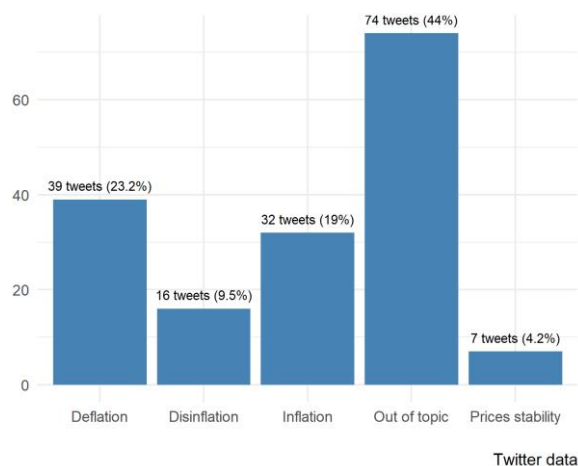


Figure 13: Description variable *what_info*

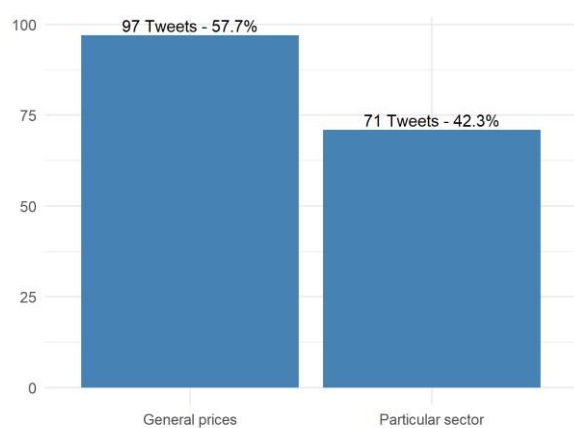


Figure 14: Description of variable *what_prices*

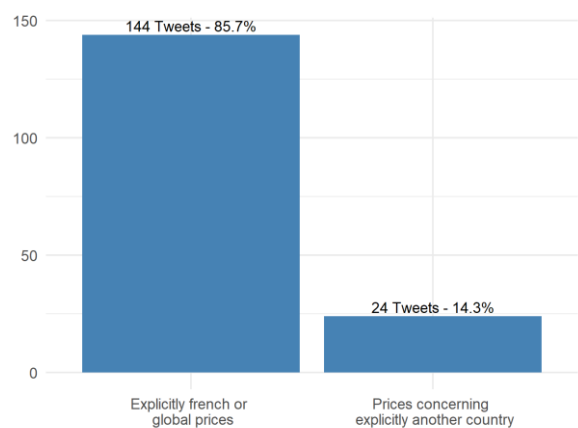


Figure 15: Description of variable *what_loc*

7.4.1. Detecting tweets related to prices problematics

To detect tweets related to prices matters, the study relies on a dictionary-based filter build on six types of lexical fields related to prices problematics. If a tweet contains one of the keywords belonging to one of these lexical fields, it is selected (first step of the methodology detailed in the main of this paper). The keywords are mainly in French, but some English words have also been included. The filter has been intentionally set broad to avoid missing any relevant tweet. However, it also captures a lot of noise that will be removed in the next steps of the methodology.

Lexical Field	Keywords
<i>Lexical field of inflation with economical terms</i>	inflation, déflation, stagflation, désinflation, inflationniste, déflationniste, antiinflationniste, antidéflationniste, ipc, ipch
<i>Lexical Field of being expensive</i>	onéreux, cher, prohibitif, coûteux, élevé, exorbitant, inabordable, conséquent, inaccessible, excessif, anormal, dispendieux, arnaque, arnaquer, ruineux, faramineux, hors de portée, rondette, inconcevable, rédhitoire

<i>Lexical field of being cheap</i>	faible, modique, avantageux, brader, imbattable, dérisoire, alléchant, réduit, occase, occasion, défiant toute concurrence, aubaine, modeste, clopinettes, bon prix, attrayant, clopinette, abordable, raisonnable, compétitif, accessible, acceptable, normaux, moyen, équitable, intéressant, convenable, négligeable.
<i>Lexical field of prices and costs</i>	prix, tarif, montant, coût, loyer, vente, achat, location, frais, abonnement, facture, coûter, facturer, payer, tarifier, vendre, devis, paiement, rabais, tarifaire, croissance, promotion, remise, ristourne.
<i>Lexical field of statistical institutions</i>	bce, banque centrale, banque central, banque de France, insee, fed, taux directeur, taux intérêt
<i>Additional keywords in English</i>	price, prices, cost, costs, rent, rents, bill, bills.

Table 9: List and content of lexical fields used to detect tweets related to prices matters

7.4.2. Detecting tweets content towards prices evolutions

In this section, we aim to specify the content of the tweet towards prices evolutions – inflation, deflation, disinflation, and prices stability – using keywords in French and in English from Table 9. For instance, if a tweet contains one of the words of the lexical field of inflation, it is classified as related to inflation. If a tweet does not contain any of the keywords included in the listed lexical fields, it is classified as “other” (out of topic). In Table 10, we summarize the heuristic rules applied, which combined lexical fields of Table 9 in this fashion.

Category	Heuristic rule
<i>Inflation</i>	[Lexical field of acceleration OR Lexical field of increase] EXCLUDING Lexical field of deflation.
<i>Deflation</i>	Lexical field of decrease OR Lexical field of deflation
<i>Disinflation</i>	[Lexical field of decrease OR Lexical field of slowdown OR “disinflation”] EXCLUDING Lexical field of deflation
<i>Prices stability</i>	Lexical field of prices AND lexical field of stabilization

Table 10: List of the keywords to specify the tweet’s content. Reading Note: a tweet is classified as talking about “inflation”, if it contains one of the keywords of the lexical field “acceleration” or one of the keywords of the lexical field “increase”, but does not contain any words belonging to the lexical field of “deflation”.

7.4.3. Keywords list to target economists and finance experts among retweeters

Each user has completed the sidebar “Description” where they describe their Twitter account content and profile. This information is available in our data and enables us to identify the users’ profile. We lists the keywords used to target the community of economists and finance experts among the Banque de France retweeters with the information available in this variable “description”, in French: *banque, actuaire, bancaire, banque, bank, assurance, finance, marche, market, investissement, bourse, business, economi, economy, credit, monnaie, entreprise, pme, tpe, eti, monetary*. If any of those words appears in their biography, then a user is considered as an economist or finance professional.

7.5. List of additional explanatory variables

This appendix describes the additional explanatory dictionary-based variables used in the Random Forest that predicts whether tweets are related to prices matters. For each tweet, each variable checks whether at least one word of a precise lexical field is found. For instance, the variable “acceleration” is set to 1 if one of the keywords related to the lexical field of acceleration is found. The variables have been built in an expert way to better characterize tweets regarding the prices problematic, according to five dimensions:

- Variables to check the presence of words related to prices matters or statistical institutions involved with the inflation problematic;
- Variables to check the presence of words related to directional evolutions to specify the prices evolution;
- Variables to check the presence of words related to the “cheap” or “expensive” lexical field to specify the perception of prices levels;
- Variables to check the presence of degree adverbs/adjective, negation terms;
- Variables to check the presence of words to be excluded (fake friends of keywords that belong to the lexical field of prices).

Most of the variables are in French, but seven variables are based on English keywords. In further developments, more variables based on English words could be added. The variables also distinguish between verb and noun in order to integrate grammatical logics into the model. Table 11 lists the additional dictionary-based variables, by describing what they attempt to capture, as well as the language and the grammatical type of their keywords. Those keywords themselves are not displayed in this paper for concision purposes.

Variable	Variable lexical field	Language	Grammatical type
1	acceleration	French	Noun
2	to accelerate	French	Verb
3	increase	French	Noun
4	to increase	French	Verb
5	decrease	French	Noun
6	to decrease	French	Verb
7	slowdown	French	Noun
8	to slow down	French	Verb
9	stabilization	French	Noun
10	to stabilize	French	Verb
11	stagnation	French	Noun
12	to stagnate	French	Verb
13	change	French	Noun
14	to change	French	Verb
15	stative verbs	French	Verb
16	expensive	French	Noun/Adjective

17	cheap	French	Noun/Adjective
18	affordable	French	Noun/Adjective
19	prices	French	Noun
20	discount	French	Noun
21	inflation (economical words)	French	Noun
22	negation terms	French	Adverb
23	little	French	Degree Adverb
24	much	French	Degree Adverb
25	little	French	Degree adjective
26	much	French	Degree adjective
27	terms to exclude	French	(no difference)
28	prices (very restrictive list)	French	(no difference)
29	prices (restrictive list)	French	(no difference)
30	prices (large list)	French	(no difference)
31	statistical institutions	French	Names
32	increase	English	Noun
33	to increase	English	Verb
34	decrease	English	Noun
35	to decrease	English	Verb
36	stabilization	English	Noun
37	to stabilize	English	Verb
38	prices	English	Noun

Table 11: List of additional dictionary-based explanatory variables

7.6. Definition of the metrics

7.6.1. Pearson and Spearman correlations coefficients

Pearson and Spearman correlation coefficients are indicators used to assess how well two variables are correlated.

The Pearson coefficient is used to measure the linear correlation between two variables. It is defined as the covariance of the two variables divided by the product of their standard deviations. It has a value between +1 and -1. A value of +1 means a positive collinearity, 0 means no linear correlation, and -1 means a negative collinearity.

Spearman's rank correlation coefficient is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It is defined as the Pearson correlation between the rank values of those two variables. While Pearson correlation assesses linear relationships, Spearman correlation assesses monotonic relationships not necessary linear. With no repeated data values, a Spearman correlation of +1 or -1 occurs when one variable is exactly a monotonic function of

the other. The Spearman correlation between two variables will be high when the observations of two variables have a similar rank, and low otherwise. Spearman coefficient is appropriate for both continuous and discrete ordinal variables.

7.6.2. Metrics for model evaluation

Our study aims to predict the value of a binary variable Y as a function of a number of explanatory variables X . More precisely, if a tweet concerns prices matters ($Y=1$) or not ($Y=0$). Supervised machine learning methods rather produce a probability than a classification. The idea is then to set a probability threshold to classify a tweet in a category. This is a classification rule and model evaluation consists in comparing predicted and true outcome values by varying this threshold. Different metrics can be used to evaluate the quality of a classification.

Confusion matrix

A classification can be qualified by a confusion matrix. It provides more insight into the performance of a predictive model by describing which classes are predicted correctly, and what types of errors are being made. In the case of a two-class classification problem, the confusion matrix is:

	Positive prediction	Negative prediction
Positive class	Count of True Positive (TP)	Count of False Negative (FN)
Negative class	Count of False Positive (FP)	Count of True Negative (TN)

Table 13: Confusion matrix

True positive is when the actual value is 1, and the predicted value is 1. For instance, the tweet concerns prices matters, and the model predicted it would. True negative is when the actual value is 0 and the predicted value is 0. For instance, the tweet does not concern prices matters, and the model predicted it would not. False positive is when the actual value is 0 and the predicted value is 1. For instance, the tweet does not concern prices matters, and the model predicted it would. False negative is when the actual value is 1 and the predicted value is 0. For instance, the tweet concerns prices matters, and the model predicted it would not.

Evaluation metrics given a probability threshold

From the confusion matrix, it is possible to calculate the following metrics to evaluate the quality of predictions by combining each of the four classes. Table 14 summarizes them, with a short explanation of their meaning.

Name	Computation	Meaning
<i>Accuracy</i>	$\frac{TP + TN}{TP + TN + FP + FN}$	What proportion of the observations are correctly classified?
<i>Precision</i>	$\frac{TP}{TP + FP}$	Among actually relevant observations, what proportion is correctly identified?
<i>Recall</i>	$\frac{TP}{TP + FN}$	Among what is identified as relevant, what proportion actually is?
<i>F-score</i>	$2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$	A harmonic combination of precision and recall

Table 14: Most standard model evaluation metrics

The most common metric used is the accuracy, because it is the most straightforward to understand. However, it can overestimate the performance of a model in the case of imbalanced data. Then, the recall and precision metrics become handy and need to be closely analyzed.

Receiver Operating Characteristic curve and Area Under the Curve

The ROC curve (or Receiver Operating Characteristic curve) is a plot that summarizes the performance of a binary classification model. Each point indicates the False Positive Rate and the True Positive Rate, for a given threshold. At (0, 0), the classifier assigns to all the observations the prediction of $Y=0$: there are no false positives, but also no true positives. At (1, 1), the classifier assigns to all the observations the prediction of $Y=1$: there are no true negatives, but also no false negatives. At (0, 1), the classifier predicts no false positives and no false negatives, and is therefore perfectly accurate. At (1, 0) the classifier has no true negatives nor true positives, and is therefore always being wrong. Simply reversing its predictions allow getting a perfectly accurate classifier. A random classifier draws a line from (0, 0) to (1, 1). The ROC curve makes it easy to compare the prediction quality of several models by simply comparing their respective ROC curves. The closer a ROC curve is from the top left corner, better is the model.

The Area Under the Curve (AUC) is an indicator of the quality of a model. The AUC values range between 0 and 1 (1 for a perfect model, 0.5 for a random model, 0 for a model always wrong). It is a great tool for model evaluation but it is not well tailored for imbalanced data because it tends to overestimate the quality of the model.

Using Twitter Data to Gauge Inflation Perception

Data Science in Central Banking: Machine Learning Applications

Julien Denes
Banque de France
19-22 October 2021

Introduction

Motivations

Why measuring inflation?

- ensuring price stability is a key role of central banks
- observing past inflation is easy, but much less interesting than future inflation
- what truly matters is how people anticipate future inflation, as they will act accordingly

Why using Twitter?

- granular data as a complement to household surveys thanks to the rich available information about the user and the tweet
- real time, high frequency data
- accessible in nearly open data and free

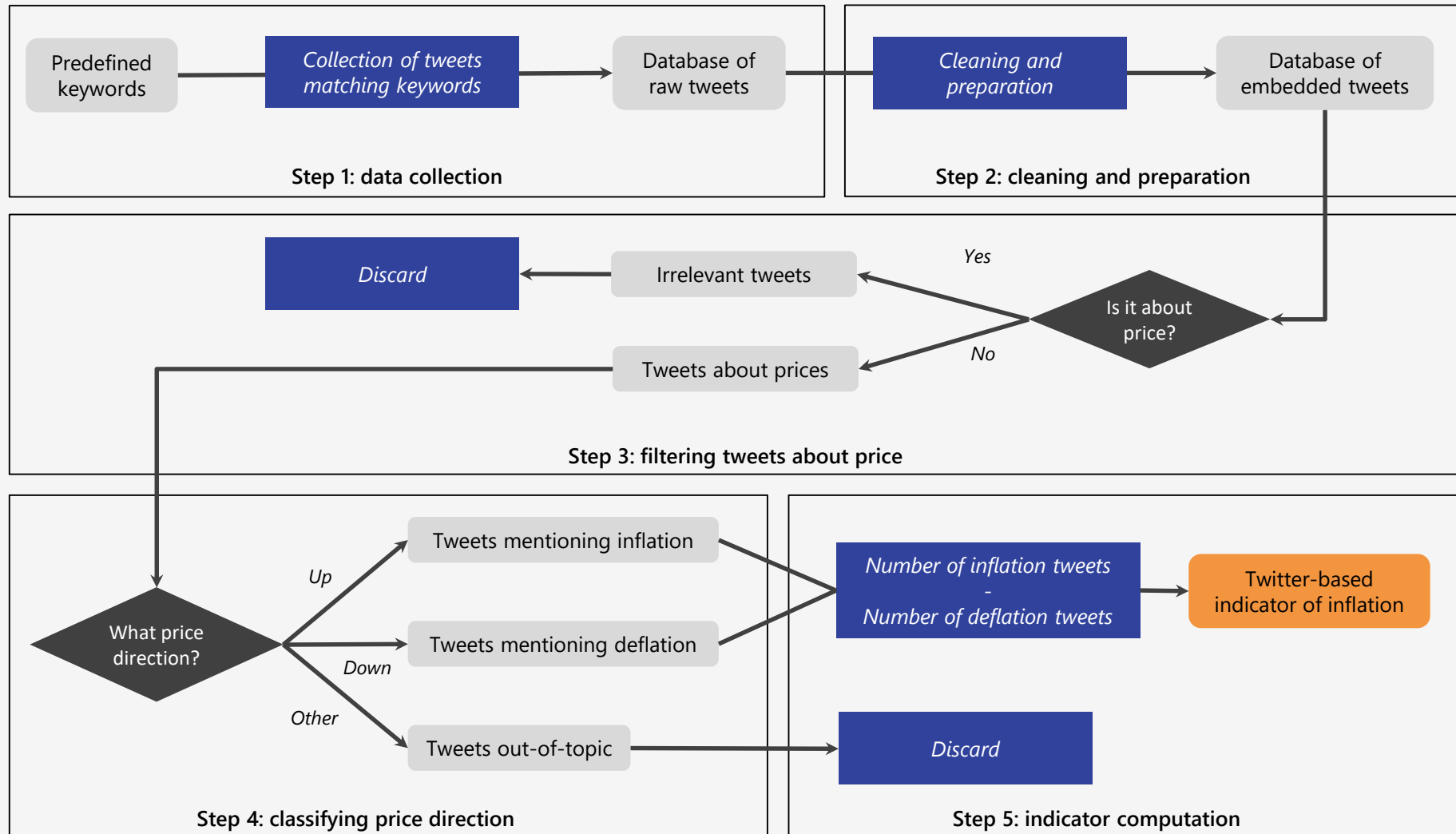
Introduction

Summary

1. Presentation of the pipeline
2. Detailed methodology of each step
3. Resulting indicators
4. Improvements and future works

Presentation of the pipeline

What do people think about the evolution of prices?



Detailed methodology

Step 1: Data collection

Keywords matching

- collecting any tweet matching a broad set of keywords, from economical terms to price-related terms and expert inflation vocabulary (≈ 100 words)

Additional filtering

- from January 2008 to June 2021, and only in French
- only a subset of users: retweeters of Banque de France's tweets ($\approx 3,500$ users)

Resulting amount of data

- more than 500,000 tweets

Detailed methodology

Step 2: Data cleaning and preparation

A standard cleaning

- removing stop words
- applying lemmatization (using only words roots) and stemming (splitting)

Text transformation using embeddings

- embedding is a representation of a text (here a tweet) as a numerical vector
- we combine word2vec embeddings and keywords-based indicators

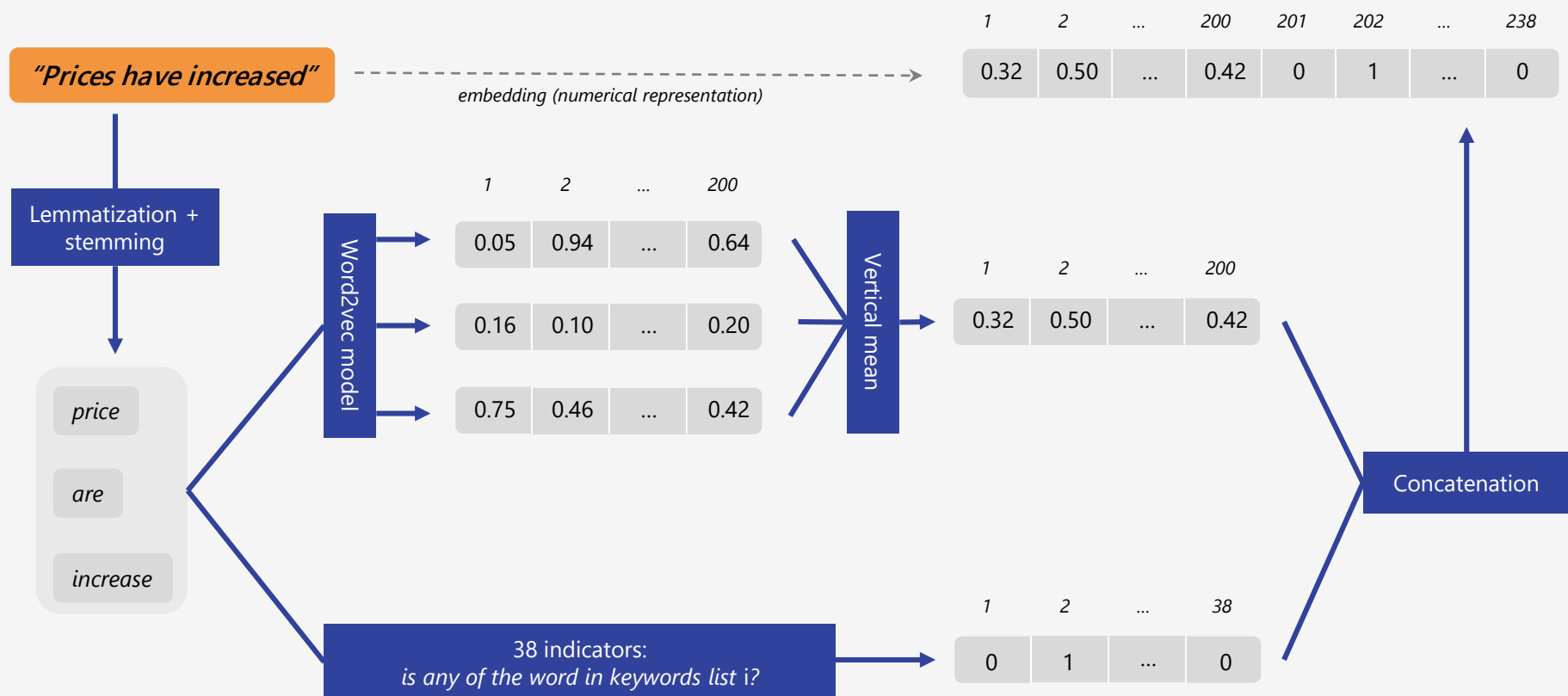
Why two types of embeddings

- word2vec provides general contextual information about the words used
- keywords indicators target the presence of specific lexical fields

Detailed methodology

Step 2: Data cleaning and preparation

An example of tweet full preparation



Detailed methodology

Step 3: Filtering tweets about price

Aim of the step

- input: a database of tweets, with their embeddings
- output: the subset of tweets that relate to price (i.e. relevant in a broad way)

Chosen model

- a random forest made of 500 trees
- major pros: fast to train and infer, light, and interpretable

Dataset used for training

- 800 tweets labelled binary as either "about price" (1) or not (0)

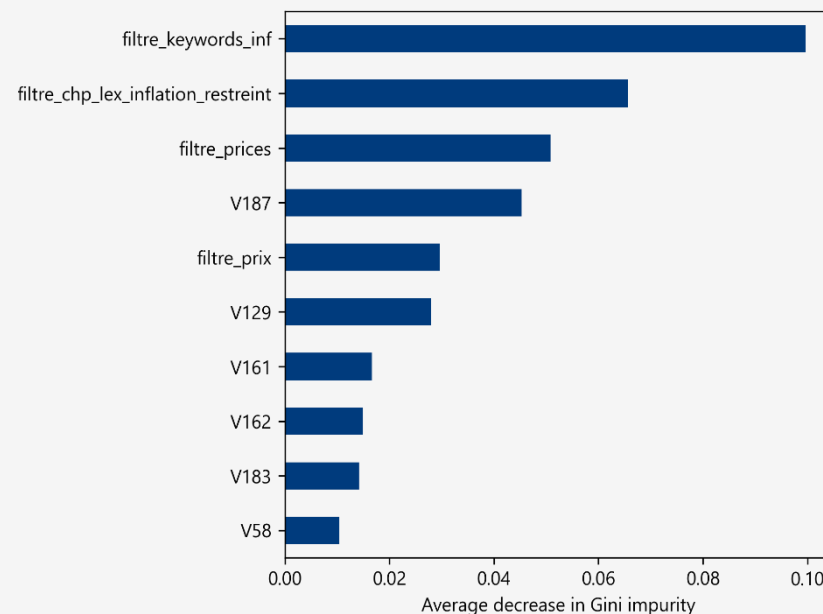
Detailed methodology

Step 3: Filtering tweets about price

Performance of the model

Metrics	Value on testing sample
<i>Accuracy</i>	90.00
<i>F1-score</i>	88.69
<i>Precision</i>	91.27
<i>Recall</i>	87.22

Feature importance



Detailed methodology

Step 4: Classifying according to price direction

Aim of the step

- input: a database of tweets about price, with their embeddings
- output: each tweet is tagged as mentioning prices going up, down, or anything else

Chosen model

- a random forest made of 500 trees (again)

Dataset used for training

- 1,100 tweets labelled as either “up”, “down”, or “other” (multi-label task)

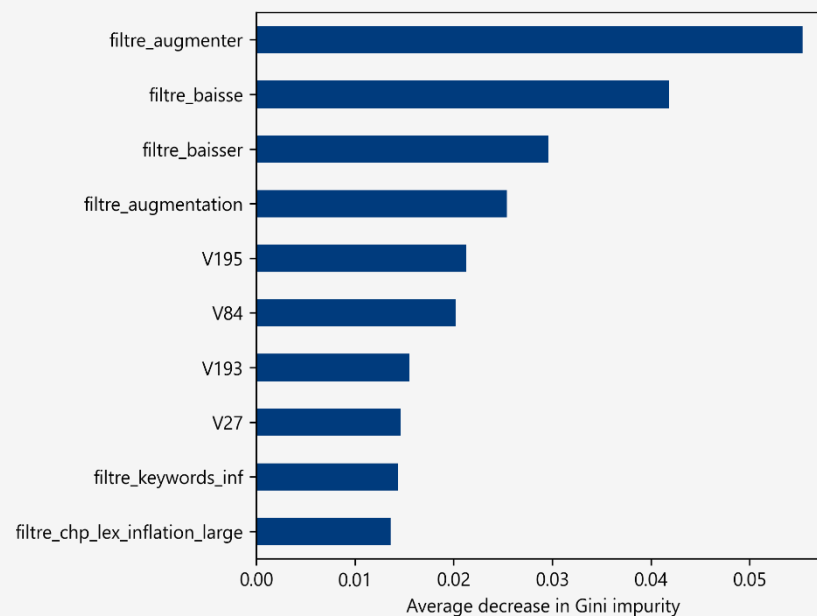
Detailed methodology

Step 4: Classifying according to price direction

Performance of the model

Metrics	Value on testing sample
<i>Accuracy</i>	85.58
<i>F1-score</i>	84.53
<i>Precision</i>	84.64
<i>Recall</i>	84.43

Feature importance



Detailed methodology

Step 5: Computing the indicator

Aim of the step

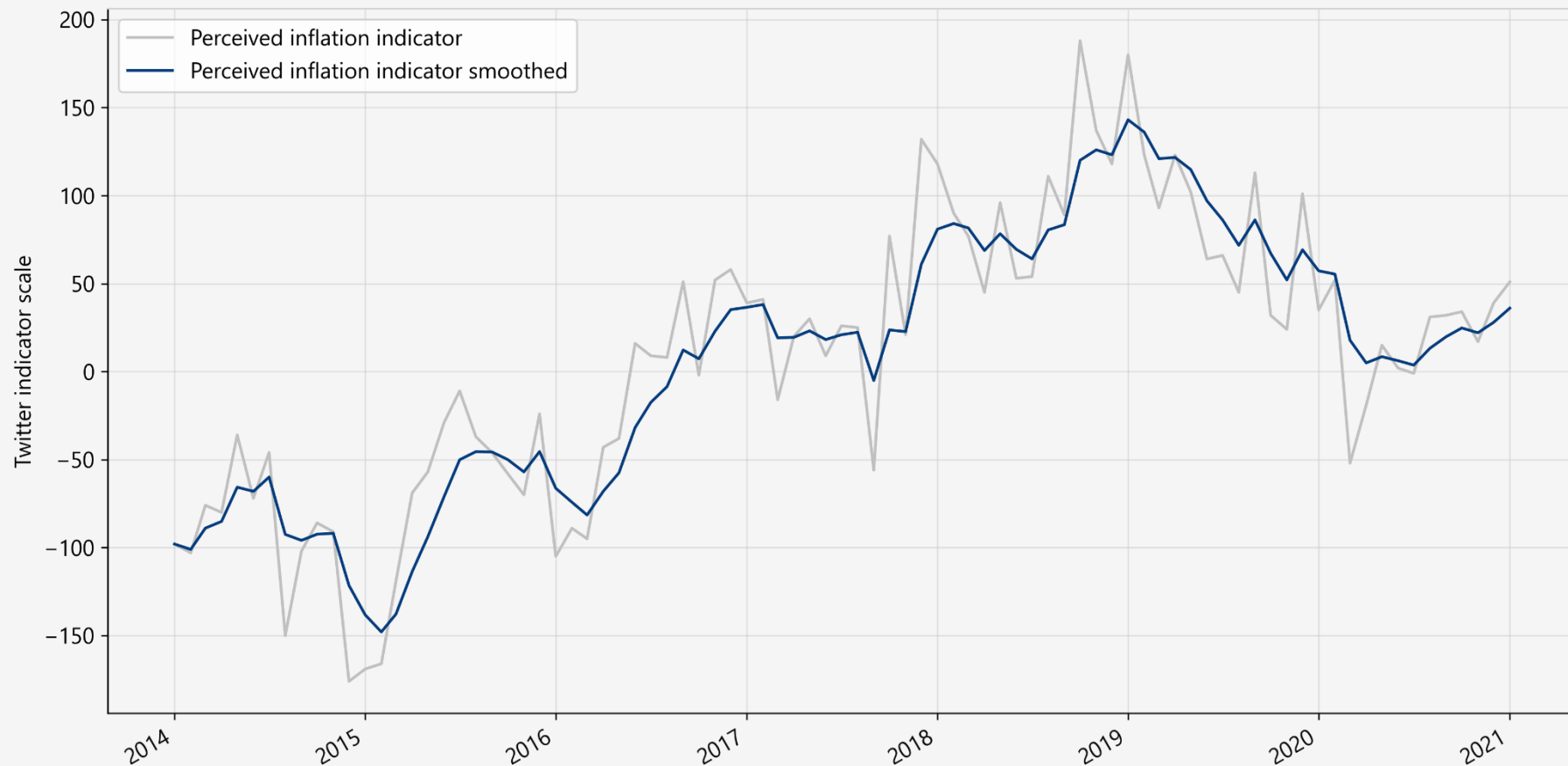
- input: a database of tweets tagged as “prices going up”, “prices going down”, or anything else
- output: an indicator of inflation as perceived by Twitter

Chosen method

- for each period of time (e.g. day, week, month), the value of the indicator is the difference between the number of tweets mentioning prices going up and those mentioning prices going down
- voluntarily simple and naïve as to mimic a balance of respondents opinion in household surveys (where answers are binary)

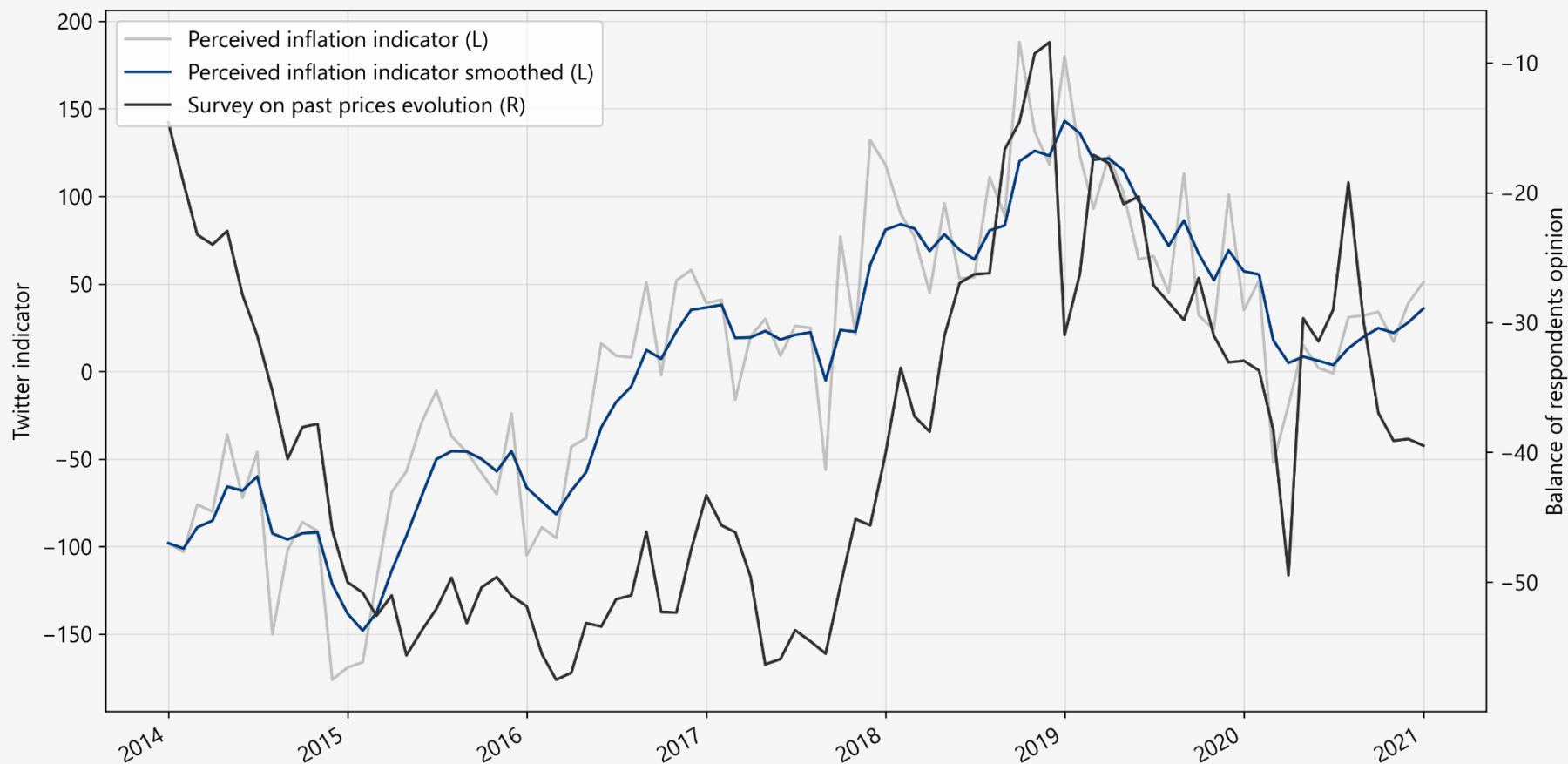
Results

Twitter indicator of perceived inflation



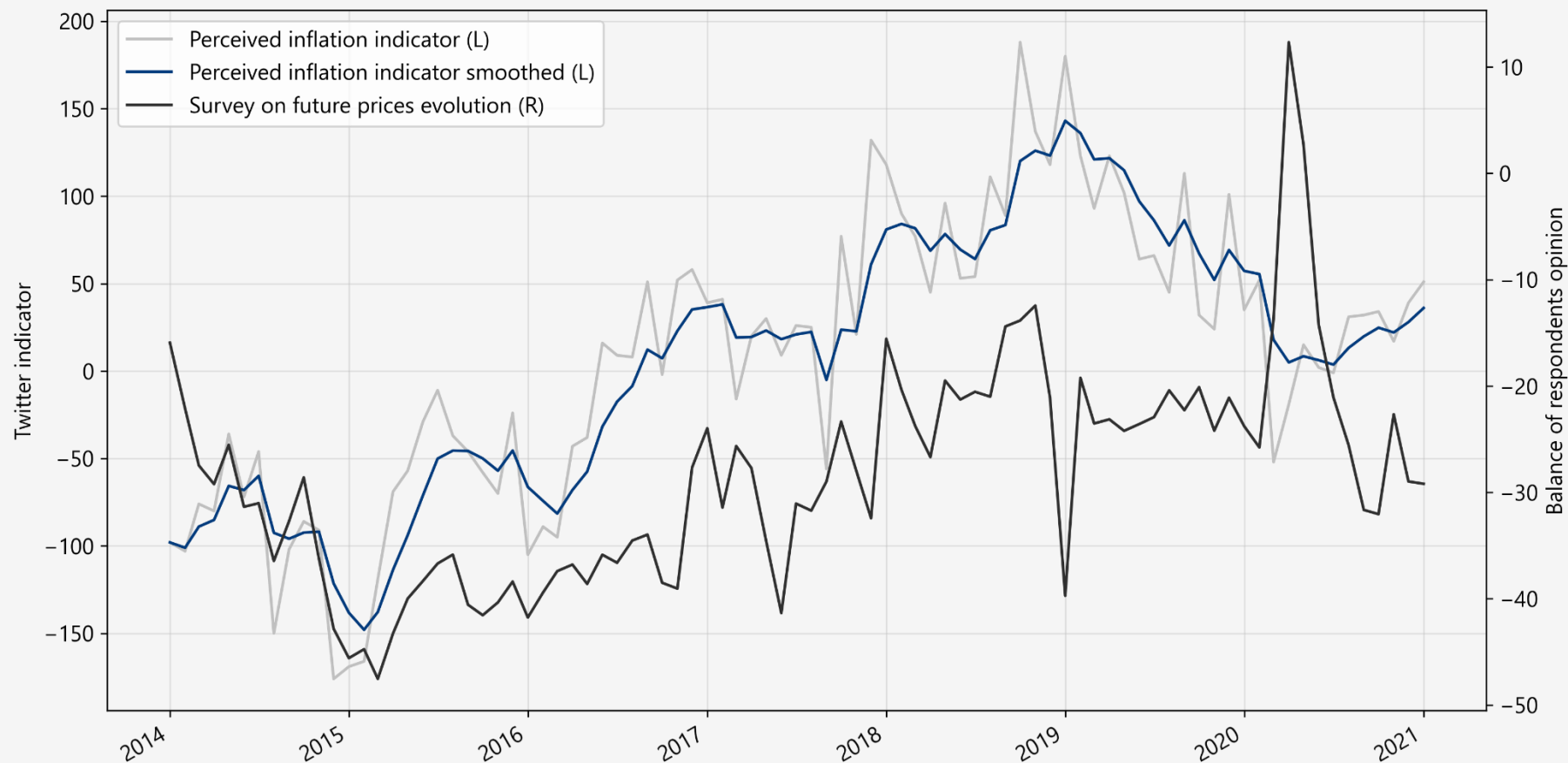
Results

Consistency with households surveys: past evolution



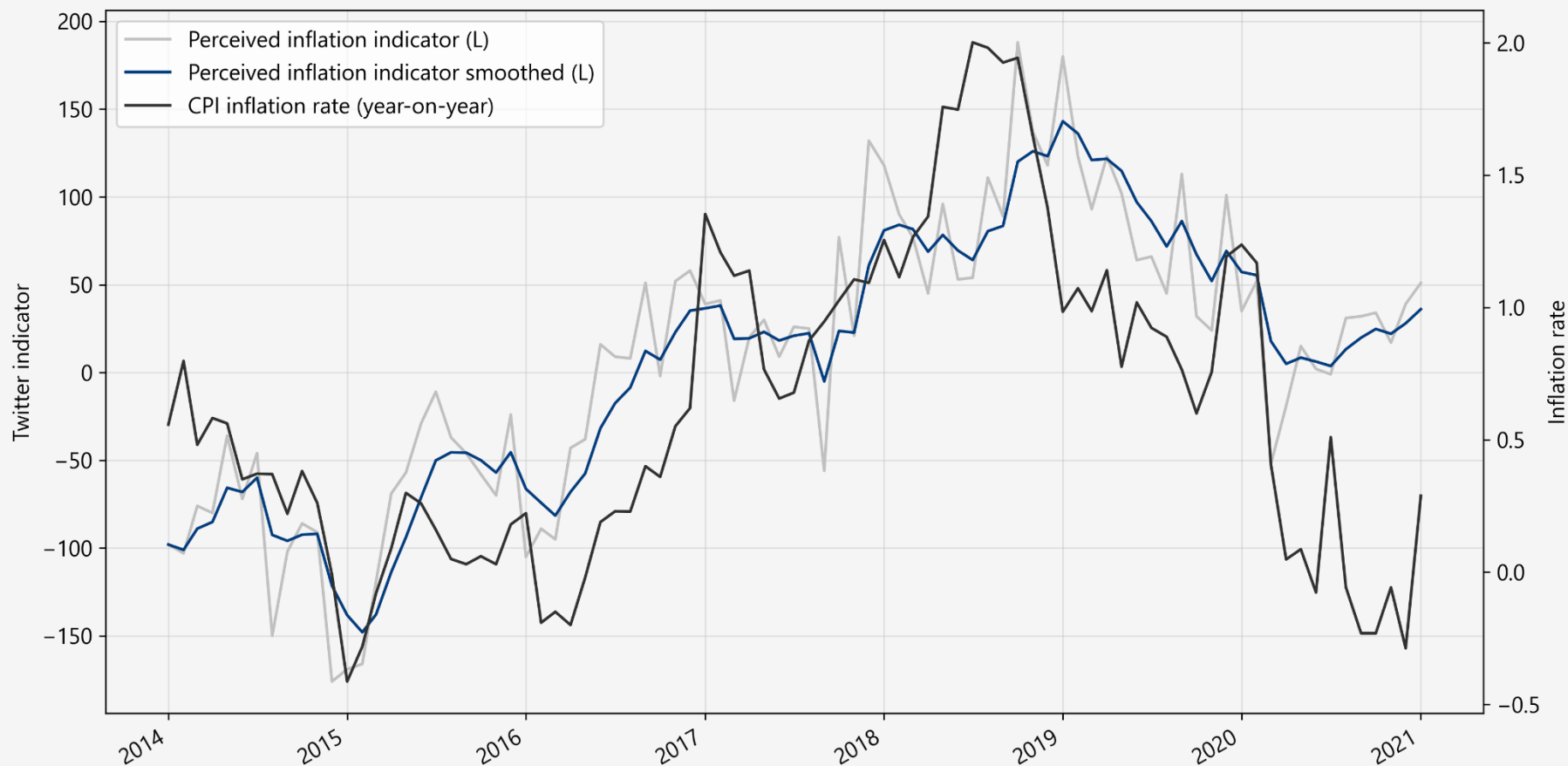
Results

Consistency with households surveys: future evolution



Results

Consistency with true inflation rate



Results

Consistency: correlation coefficients

Correlation between...	Pearson	Spearman
<i>Past evolution perception and Twitter indicator</i>	0.213	0.219
<i>Past evolution perception and smoothed Twitter indicator</i>	0.305	0.329
<i>Future evolution perception and Twitter indicator</i>	0.460	0.487
<i>Future evolution perception and smoothed Twitter indicator</i>	0.505	0.558
<i>Inflation rate and Twitter indicator</i>	0.642	0.664
<i>Inflation rate and smoothed Twitter indicator</i>	0.735	0.792

Improvements and future works

Refining data collection

- lifting the restriction on Twitter users whose tweets are collected
- potential issue: huge increase of the volume of data

Distinguishing perception and anticipation

- the true goal is to capture how people anticipate future inflation
- requires an additional step to distinguish between past/present and future

Declining topics mentioned

- eliciting what topics are mentioned, to construct sectorial indicators: consumption prices, transportation, housing, raw material, etc.

Using Twitter Data to Gauge Inflation Perception

Julien Denes

julien.denes@banque-france.fr

Direction Générale du Système d'Information (DGSi)
Direction des Données et des Services Analytiques (DDSA)
Service Intelligence Artificielle, Données, Expérimentations (ILIADE)

Appendix

Predefined keywords for tweets collection

Lexical field	Keyword
<i>Inflation (economic vocabulary)</i>	inflation, déflation, stagflation, désinflation, inflationniste, déflationniste, antiinflationniste, antidéflationniste, ipc, ipch
<i>Expensive</i>	onéreux, cher, prohibitif, couteux, élevé, exorbitant, inabordable, conséquent, inaccessible, excessif, anormal, dispendieux, arnaque, arnaquer, ruineux, faramineux, hors de portée, rondelette, inconcevable, rédhibitoire
<i>Cheap</i>	faible, modique, avantageux, brader, imbattable, dérisoire, alléchant, réduit, occase, occasion, défiant toute concurrence, aubaine, modeste, clopinettes, bon prix, attrayant, clopinette, abordable, raisonnable, compétitif, accessible, acceptable, normaux, moyen, équitable, intéressant, convenable, négligeable
<i>Prices and costs</i>	prix, tarif, montant, coût, loyer, vente, achat, location, frais, abonnement, facture, coûter, facturer, payer, tarifier, vendre, devis, paiement, rabais, tarifaire, croissance, promotion, remise, ristourne
<i>Statistical institutions</i>	bce, banque centrale, banque central, banque de france, insee, fed, taux directeur, taux intérêt

Appendix

Lexical fields used for embeddings indicators

Variable	Variable lexical field	Language	Grammatical type
1	acceleration	French	Noun
2	to accelerate	French	Verb
3	increase	French	Noun
4	to increase	French	Verb
5	decrease	French	Noun
6	to decrease	French	Verb
7	slowdown	French	Noun
8	to slow down	French	Verb
9	stabilization	French	Noun
10	to stabilize	French	Verb
11	stagnation	French	Noun
12	to stagnate	French	Verb
13	change	French	Noun
14	to change	French	Verb
15	stative verbs	French	Verb
16	expensive	French	Noun/Adjective
17	cheap	French	Noun/Adjective
18	affordable	French	Noun/Adjective
19	prices	French	Noun
20	discount	French	Noun
21	inflation (economical words)	French	Noun
22	negation terms	French	Adverb
23	little	French	Degree Adverb
24	much	French	Degree Adverb
25	little	French	Degree adjective
26	much	French	Degree adjective
27	terms to exclude	French	(no difference)
28	prices (very restrictive list)	French	(no difference)
29	prices (restrictive list)	French	(no difference)
30	prices (large list)	French	(no difference)
31	statistical institutions	French	Names
32	increase	English	Noun
33	to increase	English	Verb
34	decrease	English	Noun
35	to decrease	English	Verb
36	stabilization	English	Noun
37	to stabilize	English	Verb
38	prices	English	Noun

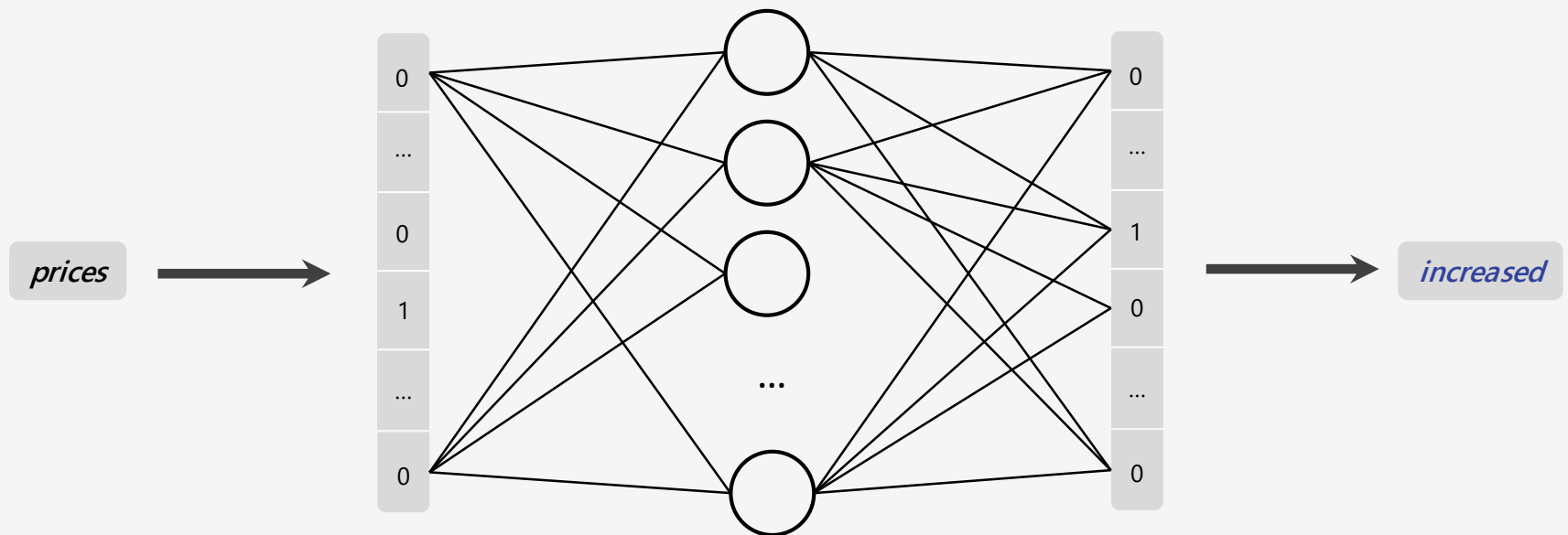
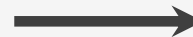
Appendix

Focus on word2vec: how it works

"I think that prices have increased quite fast this year"

*"I think that prices have **increased** quite fast this year"*

(prices, **increased**)
(have, **increased**)
(quite, **increased**)
(fast, **increased**)



Input

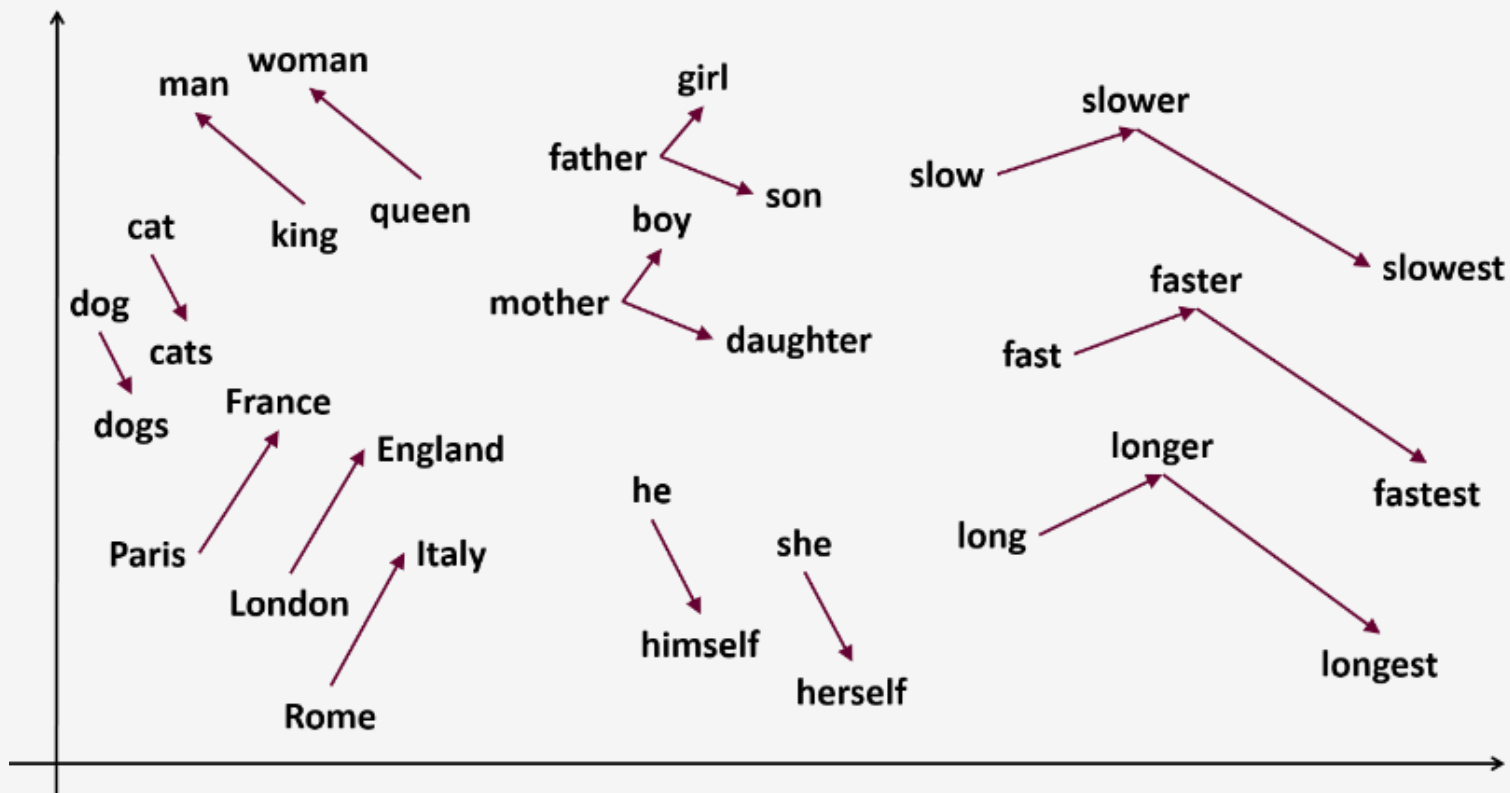
Input as one-hot-encoding

Model: one layer of 200 neurons

Output to predict

Appendix

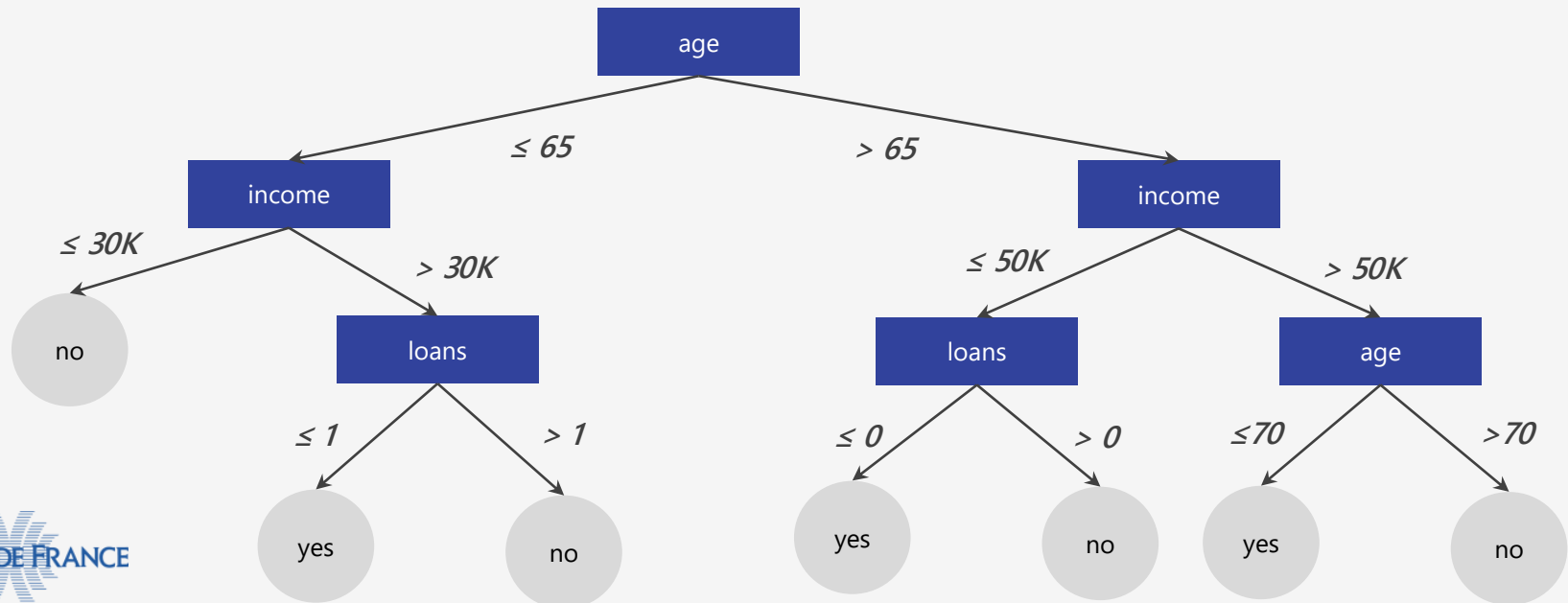
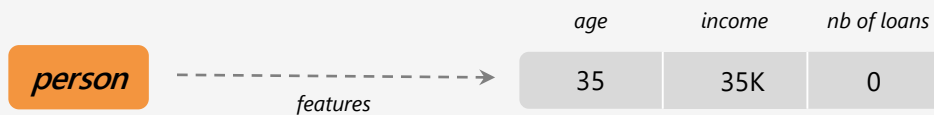
Focus on word2vec: properties



Appendix

Focus on random forests: decision tree

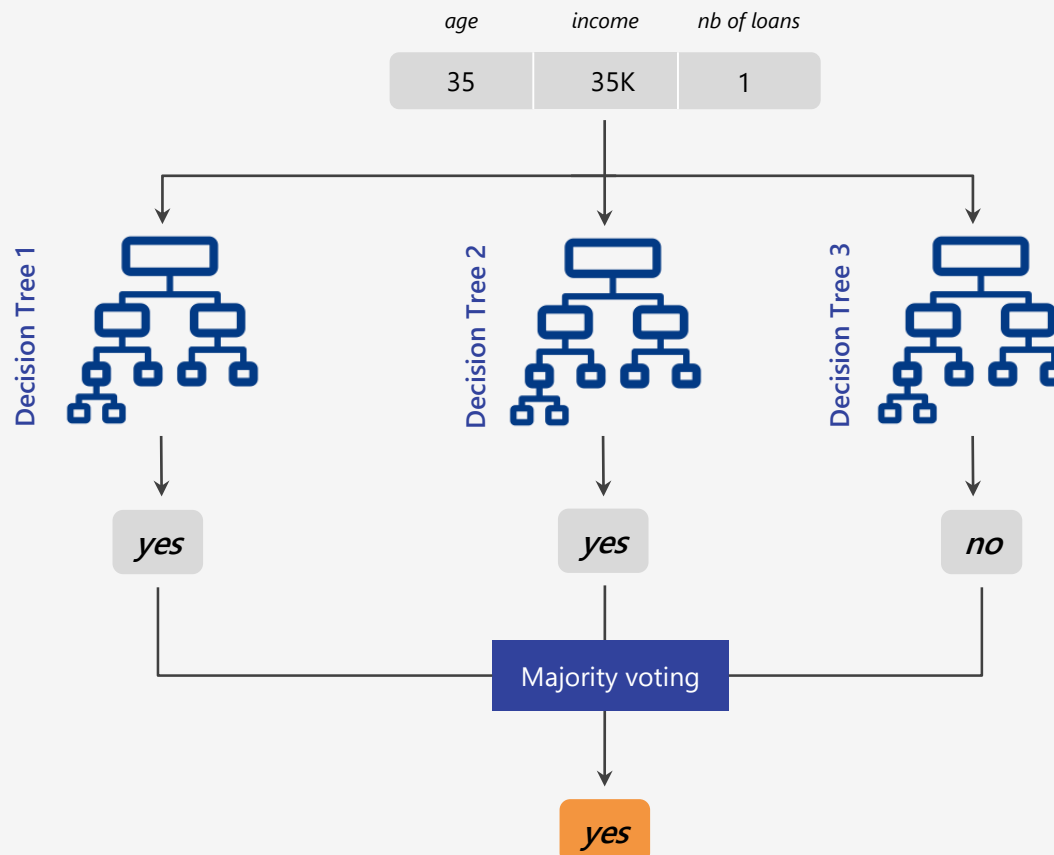
Will a person get a loan from their bank?



Appendix

Focus on random forests: forests

Will a person get a loan from their bank?



Appendix

Sources and references

Seminal work of Banca d'Italia


- Angelico C., Marcucci J., Miccoli M. and Quarta F. (2021). [Can we measure inflation expectations using Twitter?](#) *Temì di discussione* n° 1318.

INSEE monthly household surveys

- [Series on opinions about prices evolution.](#)

INSEE consumption price index, base 2015

- [Series and documentation.](#) Inflation rate is computed by us as the 12-month relative difference.



This presentation describes preliminary results of ongoing research. It is made available to the public solely to elicit discussion and comments. Views expressed in the presentation are those of the authors and do not reflect the position of the Banque de France.