The similarity of ECB's communication, Replication

Grahl, Lukas; Thomas, Joel

November 24, 2023

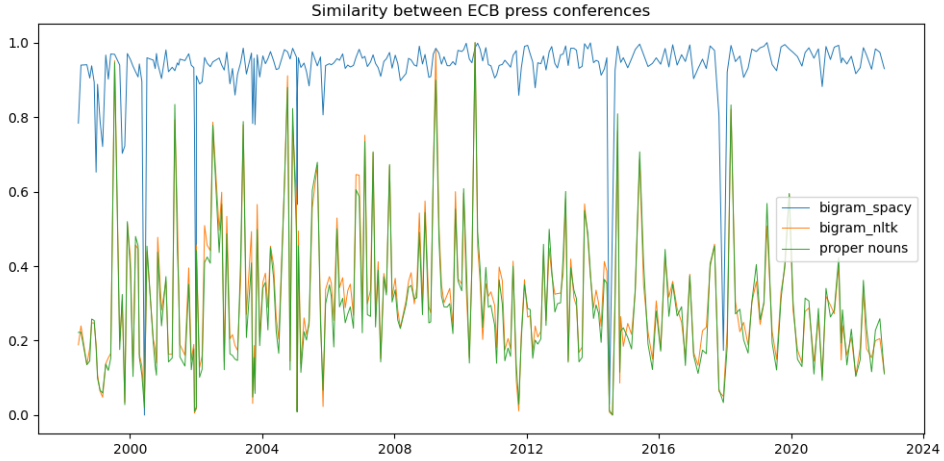
1 Data collection

In collection the ECB press releases this paper depart from **pc_link.xlsx**. In a first steps two links, one linking the speech overview page, the other containing no press conference but merely questions and answers are removed. For the remaining link the webpages for each press release are scraped from the ECB website using BeautifulSoup.
Once the HTML document for each press release is obtained the text, title and publication date are extracted. Both title and date are easily accessible and require no further logic. The press release subtitle is more intricate but identifiable for all press releases by its *class*. The speech text presents some further difficulty. For the majority of press releases the speech is contained in a *¡div¿* object together with the press release subtitle. This makes it easy to separate speech and question and answer. Especially older text releases do not follow this structure, instead an *¡a¿* link to the question and answer section is contained in the same *¡div¿* object. Some press releases have even different structure, for which we identify the speech as ending on an invitation to ask questions. While this identifier would most likely apply to most speeches we prefer the two above ones, as they provide unique identification. This makes the code less error prone and allows one to notice possible future changes in the HTML structure.
Having obtained the text of each press release we follow (ref) in removing stop words. Stop words are words carrying little or no meaning in a sentence. As such they disturb measures of similarity, making documents appear more similar than they are. It is therefore advisable to remove stop words prior to similarity analysis. However, stop words are also context sensitive, what is meaningless in one context might be important in another. Especially when expressions are carefully chosen or attenuated, as with ECB press conferences, a word such as 'serious' proceeding 'economic crisis' dramatically changes the context. For that reason we have checked the standard stop words contained in Spacy an decided to remove the below:

call, amount, against, above, up, down, serious, quite

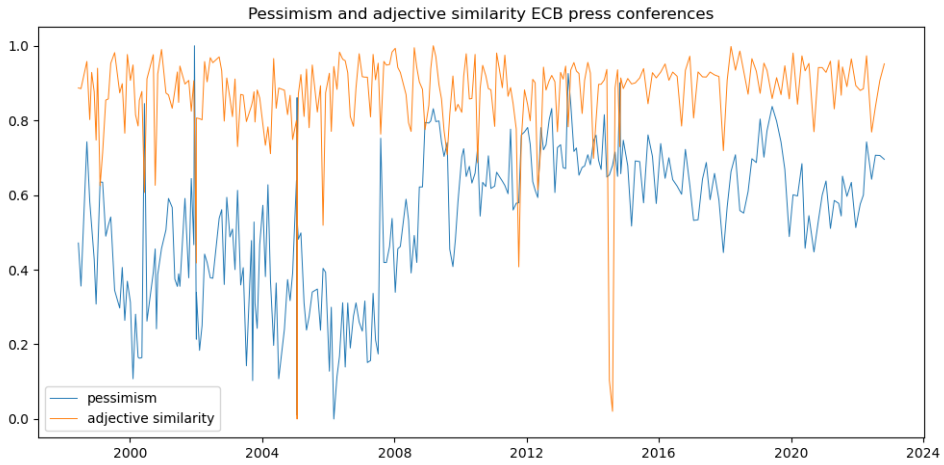
Once stop words have been removed we proceed with calculating the bigram based Jaccard similarity (ref). However, as bigrams contain all non-stop words in the text they are a rather broad representation of similarity. We therefore supplement the analysis by calculating similarity scores only based on adjectives and and proper nouns. These depart

from the motivation adjectives in ECB press releases are chosen with particular attention. As such they likely reveal much about the current state of the economy. Proper nouns are all expression describing physical things. They reflect the subject of the press releases, if for example a new topic such as a war was to arise this should be reflected across proper nouns. One could argue that named entities might be more suitable for this purpose, as they are more restrictive and trained to only capture entities. However, as named entity recognition is model based it often fails to detect unseen entities such as Covid-19 early on.
Having rescaled the data for plotting purpose one can see the evolution of similarity over time. As with the paper we can see some increase in NLTK bigram similarity over time especially after the 2008 financial crisis. However, the surge is not nearly as stark as portrayed in the paper (page 240). This is surprising given that b_nltk follows the procedure outlined by the paper. Interestingly, the spacy proper noun similarity *propns* closely traces nltk based bigrams. This offers an interesting perspective on bigrams, showing that this score is primarily driven by proper nouns, that is entities in a wider sense. However, this underscores the need for an analysis of adjectives in order to capture the urgency of the ECB press releases. The paper did so using the Loughran and McdDnald dictionary (ref).



The below traces pessimism across the ECB press releases as proivded by the below identity.

$$P_t = \frac{pessimistic_t - optimistic_t}{TotalWords_t}$$



As in our reference paper excess returns are defined as a 201 days rolling mean returns up to 50 days prior of t.

$$barR_{it} = frac1201sum_{i=-250}^{-50}R_{it}$$

The cumulative abnormal return is then specified as the the sum of returns across five days prior, the anoucement day and five days after the anoucement.

$$CAR_i = sum_{t=-5}^{5}R_{it} - barR_{it}$$