Soft and Hard Criticism in EU Progress Reports

Samuel Fraley, Olta Reçica, Timothy Moynihan

March 31, 2026

Group: TOS

Introduction to Text Mining and Natural Language Processing
Data Science for Decision Making / Data Science Methodology
Barcelona School of Economics

Abstract

This paper uses a dictionary-based NLP approach to measure hard and soft criticism in EU Progress Reports across ten candidate countries from 2019 to 2025. We find that judiciary and anti-corruption paragraphs consistently attract the harshest language while economic paragraphs show no significant severity effect, that country differences in tone persist after controlling for topic composition and time, and that there is no systematic shift in criticism after 2022.

Instructor: Prof. Dr. Hannes Mueller

1 Introduction

In 2022, Russia's invasion of Ukraine changed the logic of European Union enlargement (Dimitrova, 2024). What had been a slow, technical process of institutional alignment suddenly turned into a question of regional security and political stability (Dimitrova, 2024). For some countries, the effects were immediate and visible: Bosnia and Herzegovina went from stalled candidate to active negotiator in under two years. Ukraine and Moldova received candidate status within months while Georgia followed. Countries that had been waiting for over a decade started moving, while others remained stalled or in progress. This shift happened for reasons that have little to do with the pace of their judicial reforms or anti-corruption progress. Historically, as Gancia et al. (2019) lay out, economic unions of countries that both are of different sizes economically and have different values are less likely to succeed. But this raises a deeper question: If geopolitical logic was strong enough to reshape enlargement so visibly after 2022, was it already quietly shaping how the EU wrote about candidate countries before that?

We focus on the six countries with continuous coverage from 2019 and ask if the severity of EU criticism in progress reports vary across countries and policy domains in ways that go beyond what topic composition and time trends alone can explain. Including the fast-tracked countries in any systematic analysis would muddy the picture because they are geopolitical outliers by design. Table 1 clearly shows how fast accession statuses changed post-war compared to the years before.

	Core Sample (2019–2025)						Post-2022 Additions			
Year	**Montenegro**	**Serbia**	**Albania**	**N. Macedonia**	**Turkey**	**Kosovo**	**Bosnia & H.**	**Ukraine**	**Moldova**	**Georgia**
2019	Negotiating	Negotiating	Candidate	Candidate	Frozen	Potential	Potential	–	–	–
2020	Negotiating	Negotiating	Candidate	Candidate	Frozen	Potential	Potential	–	–	–
2021	Negotiating	Negotiating	Candidate	Candidate	Frozen	Potential	Potential	–	–	–
2022	Negotiating	Negotiating	Negotiating	Negotiating	Frozen	Applied	Potential	Candidate	Candidate	Potential
2023	Negotiating	Negotiating	Negotiating	Negotiating	Frozen	Applied	Candidate	Negotiating	Candidate	Candidate
2024	Negotiating	Negotiating	Negotiating	Negotiating	Frozen	Applied	Candidate	Negotiating	Negotiating	Halted
2025	Negotiating	Negotiating	Negotiating	Negotiating	Frozen	Applied	Candidate	Negotiating	Negotiating	Halted

Table 1: EU Accession Status by Country and Year

This gives us a balanced panel of candidates where variation in EU language is harder to explain by a single external shock. These are the standard, the ones the EU has been writing about for years under a more or less consistent framework, and strategic considerations shape the language even in these cases. Using a dictionary-based NLP pipeline applied to roughly 26,800 paragraphs across ten countries, we construct measures of hard and soft criticism and a severity ratio that captures how sharp the language is in each report. Even among the pre-2022 candidates, the variation is striking. Turkey receives by far the harshest language in the sample, with hard criticism running consistently above every other country. Serbia sits at the opposite end, receiving reports dense with concern but sparse in outright condemnation, even as its reform record remains contested.

The geopolitical logic the EU made explicit in 2022 was already quietly operating in how it wrote about candidate countries before that. The war did not invent strategic calibration in EU language, it just made it impossible to ignore. Our findings are a window into how that calibration worked before it became official policy. To capture this, we tag paragraphs across four policy domains: judiciary, corruption, governance, and economy. We next estimate logistic and OLS regression models on the six countries with complete panel coverage, controlling for topic composition,

year trends, and country fixed effects. We find that criticism severity varies systematically across both dimensions. Judiciary and anti-corruption paragraphs concentrate the hardest language while economic paragraphs show no significant effect, consistent with the idea that legal convergence is where the EU draws its hardest lines. Country differences persist after these controls, with Turkey attracting significantly harder language and Serbia significantly softer, a pattern that is difficult to explain by topic composition alone.

2 Data

EU Progress Enlargement Candidate Reports (Enlargement and Eastern Neighbourhood Department of the EU, 2026) were scraped as plain text across 10 candidate and potential candidate countries which are Albania, Bosnia and Herzegovina, Georgia, Kosovo, Moldova, Montenegro, North Macedonia, Serbia, Turkey, and Ukraine, covering report years 2019-2025, though not all countries have full temporal coverage. After segmentation and cleaning, this yields 26,768 paragraph-level observations. Each paragraph is the unit of analysis, stored alongside its country, year, paragraph ID, and word count.

Table 2: Paragraph-level data structure

Field	Data Type	Description
Paragraph Text	Document	Paragraphs between 50 and 500 words drawn from each country's annual report
Country	Metadata	The country represented as plain text
Year	Metadata	The year of the report
Paragraph ID	Metadata	Unique identifier for each paragraph within a country-year, starting at 0
Word Count	Metadata	Total number of words in the paragraph, stored as an integer

Note: Document-type fields contain the raw text used for scoring; metadata fields are used for grouping and identification only.

3 Methods

3.1 Key Design Choices

3.1.1 EPU-Style Dictionary Scoring

Following Baker et al. (2016), paragraphs are scored by counting hits from a custom keyword dictionary for both hard and soft criticism, normalized per 1,000 words. This approach was chosen over machine learning alternatives for three reasons: (1) interpretability: every score is traceable to specific words; (2) no labeled training data is required; (3) the EU's institutional register is stable and rule-governed, making dictionary methods reliable. Morphological variants are handled automatically by stemming rather than manual enumeration, keeping the dictionary compact without sacrificing recall.

3.1.2 Criticism Dictionary

A custom two-tier lexicon was built to distinguish the EU's rhetorical registers. Construction followed an iterative seeding process: we began with manual seed terms drawn from common EU critical language (*failed*, *violation*, and *backsliding* for hard criticism; *concerns*, *risk*, and *issues* for soft) then expanded each list by inspecting already-flagged paragraphs and identifying recurring similar terms. Because single-word terms are stemmed before matching, the expansion focused on capturing genuinely distinct critical concepts rather than inflectional variants of existing seeds. The example lists are below with the full lists in supporting files:

- Hard criticism (CRITICISM_HARD) — unambiguous failure language signaling explicit condemnation: failed, backsliding, violation, state capture, obstruction, impunity, deterioration, stagnation, lack of, shortcomings, inadequate, insufficient
- Soft criticism (CRITICISM_SOFT) — hedged concern language common in EU bureaucratic text: concerns, challenges, delays, limited, persistent, remain, issues, risks, needs to be, has yet to

Note: The full list of the criticism dictionary is limited to unique values after trimming to ensure that double-counting does not occur.

Across the corpus, 9,125 paragraphs (38%) contained hard criticism and 17,964 (67%) contained soft criticism. The 68% soft coverage reflects a deliberate trade-off: broader terms like *remains* and *requires* risk capturing neutral language, but the coverage level, neither too low nor too high, suggests the net is wide without being indiscriminate.

The severity ratio (hard p1k / soft p1k) captures the balance between the two registers, where a high ratio indicates that the EU is condemning rather than merely flagging concerns. This ratio is motivated by the structure of EU institutional prose, which combines critical and hedging language in every report. Raw hard criticism counts therefore reflect both genuine severity and overall report length and topic. The ratio between hard and soft criticism controls for this baseline hedging, capturing how concentrated the condemnatory language is relative to the diplomatic register that always surrounds it. To confirm that expanded terms captured genuine rhetorical neighbors rather than arbitrary additions, we audited corpus-wide term frequencies and verified that expanded terms co-occurred heavily in seed-flagged paragraphs.

3.2 Preprocessing and Pipeline

3.2.1 Corpus construction

Raw report texts are stored as plain-text .txt files whose filenames encode country and year (e.g. Albania_2019_raw.txt). A regex pattern extracts this metadata at load time, attaching country and year labels to every paragraph derived from that file. Each document then undergoes a two-pass cleaning procedure before segmentation. The first pass strips page-break markers and normalizes irregular line endings introduced by the scraper. The second pass splits the cleaned text on double newlines and on sentence boundaries preceding a capitalised word, recovering paragraph structure that would otherwise be lost at page transitions. This yields a set of candidate paragraphs per document. A minimum of 50 and a maximum of 500 words per paragraph is then applied to exclude

outliers. Paragraphs under 50 words were frequently footnotes, headers, or other formatting artifacts rather than substantive report text. Paragraphs exceeding 500 words were split at sentence boundaries into sub-units of no more than 500 words, ensuring that scoring windows remain comparable across documents. This process yielded a total of 26,768 paragraphs across ten countries and report years 2019–2025.

3.2.2 Stemming and Scoring

Before scoring, single-word terms in both dictionaries are stemmed using the Snowball stemmer (Porter2 algorithm) via NLTK. Multi-word phrases (e.g. *lack of*, *has not been implemented*) are matched against the original lowercased text without stemming, since stemming boundary effects can corrupt phrase boundaries. Single-word terms are matched against a pre-stemmed version of each paragraph, where every token has been passed through the same stemmer. This ensures that morphological variants not enumerated in the dictionary (e.g. *deteriorating* alongside *deteriorated*) are captured without manual expansion of the term lists.

Counts are normalized per 1,000 words of the original (unstemmed) paragraph to keep the denominator consistent with the word-count metadata. The severity ratio is undefined and set to `NaN` when no soft criticism is present. These three metrics form the core dependent variables of the analysis.

3.2.3 Topic Tagging

Topic areas were selected by examining the chapter structure and thematic clusters used in the EU Progress Reports themselves, grounded in the literature's distinction between structural and economic reform dimensions. This yielded four domains: judiciary, corruption, governance, and economy, with governance absorbing both public administration and political functioning from an earlier four-topic scheme. Construction followed the same seeding approach as the criticism dictionary: seed terms were drawn from report tables of contents and section headings, flagged paragraphs were inspected, and term lists were expanded iteratively:

- Judiciary: *judge, court, rule of law, justice, prosecution, tribunal, appeal, verdict*
- Governance: *governance, ministry, regulation, legislative, parliament, decentralization*
- Corruption: *corruption, fraud, embezzlement, bribery, integrity*
- Economy: *economic, market, budget, unemployment, inflation, investment, fiscal*

Topic matching uses substring search on the original lowercased text without stemming, since the topic keywords are short and distinctive enough that morphological variation is not a material concern. 22,786 paragraphs (85% of the corpus) matched at least one topic, with a mean of 1.45 topics per paragraph, solid coverage given that the full report text was analyzed with minimal exclusions.

3.2.4 Aggregation

Paragraph-level scores are averaged up to the country-year level, producing a summary table of mean hard criticism, soft criticism, and severity ratio for each country-year combination. This aggregated table is the basis for the trend plots and choropleth maps presented in the results. Paragraph scores are also aggregated separately by country, year, and topic domain to support the topic-level descriptives reported in the analysis.

3.2.5 Robustness to Preprocessing Choices

Country-level scores are insensitive to the upper word limit: replacing the 500-word cap with 300 words shifts no country's mean hard criticism score by more than +/-0.3%. Lowering the minimum from 50 to 30 words produces larger changes, up to 4.7% for Albania, because the additional short paragraphs tend to be transitional or descriptive rather than evaluative. The country ranking is nevertheless identical under both alternative specifications, indicating that the core cross-country comparisons are not an artifact of the paragraph length thresholds chosen. Switching from stemmed to unstemmed matching compresses hard-criticism coverage from 38% to 17% of paragraphs, but leaves country rankings unchanged, confirming that stemming affects overall score levels symmetrically across countries rather than introducing differential bias between them.

3.3 Analysis

Paragraph-level scores are aggregated to the country-year level by taking the unweighted mean of each metric across all paragraphs within a country-year. Per-1,000 word normalization already accounts for paragraph length, so additional word-count weighting would be redundant. This produces 57 country-year observations, one per country-year combination. We first track all three scores over time for the six countries with full temporal coverage (2019–2025), allowing us to assess whether criticism levels have shifted across the reporting period and whether differences between countries are persistent or episodic.

To isolate country-level effects from topic composition, we estimate three paragraph-level models on the six full-coverage countries (19,914 paragraphs), with North Macedonia as the reference category and standard errors clustered by country. M1 and M2 are logistic regressions; M3 is OLS estimated on paragraphs with soft criticism present, since the severity ratio is undefined when soft criticism equals zero. All models include binary topic indicators and a median-centred year trend alongside country fixed effects, allowing us to separate what the EU is writing about from how critically it is writing about each country.

4 Results

	M1 hard flag (OR)	M2 soft flag (OR)	M3 severity ratio (β)
Country fixed effects (ref: North Macedonia)			
Albania	0.746***	0.940***	−0.075***
Kosovo	1.096***	1.133***	0.011***
Montenegro	0.844***	0.773***	−0.015***
Serbia	0.829***	1.130***	−0.089***
Türkiye	1.274***	1.027***	0.037***
Topic indicators			
Judiciary	1.319***	1.055	0.088***
Corruption	1.453***	1.342***	0.071***
Governance	1.575***	1.739***	0.042*
Economy	1.017	1.077*	−0.015
Time			
Year trend	0.994	1.021	0.001
n	19,914	19,914	17,834

Note: * $p < .05$ ** $p < .01$ *** $p < .001$. Standard errors clustered by country. M1 and M2 are logistic regressions; M3 is OLS estimated on paragraphs with soft criticism present.

Table 3: Paragraph-level regression results

The three models are estimated on 19,914 paragraphs across the six full-coverage countries, with North Macedonia as the reference and standard errors clustered by country. We report odds ratios (ORs) for M1 and M2 and OLS coefficients for M3.

4.1 Topic effects

Gancia et al. argue that integration friction concentrates in legal and institutional domains rather than economic ones. Our results support this. Judiciary paragraphs are the most reliably associated with hard language: controlling for country and year, they are 32% more likely to contain hard criticism than untagged paragraphs (OR = 1.32, $p < 0.001$) and carry the largest positive shift in the severity ratio ($\beta = 0.088$, $p < 0.001$). The corruption paragraphs follow closely, with a 45% higher odds of hard criticism (OR 1.45, $p < 0.001$) and a significant severity uplift ($\beta = 0.071$, $p < 0.001$). Governance paragraphs are associated with higher soft coverage (OR 1.74, p ¡ 0.001) and a small but statistically significant increase in the severity ratio, suggesting the EU flags institutional concerns broadly while only modestly escalating to harder criticism. Economy shows a not significant association with hard criticism (OR 1.02) and carries no significant severity effect ($\beta = -0.015$). The small positive correlation for soft criticism (OR = 1.08, $p < 0.05$) suggests that economic paragraphs attract more critical soft language in absolute terms but without much change to the ratio. This topic hierarchy holds across all three dependent variables.

4.2 Country effects

Controlling for topic and year does not explain away the cross-country differences, which suggests the EU's tone reflects something beyond each country's reform record. Türkiye attracts the hardest language in the sample: paragraphs about Türkiye are 27% more likely to contain hard criticism than equivalent North Macedonia paragraphs (OR = 1.27, $p < 0.001$), with the highest severity ratio shift in the sample ($\beta = +0.037$). Kosovo also sits significantly above the reference on hard criticism (OR = 1.10, $p < 0.001$). Serbia presents the sharpest contrast: it has below-average odds of hard criticism (OR = 0.83, $p < 0.001$) yet above-average soft coverage (OR = 1.13, $p < 0.001$) and the most pronounced divergence between soft and hard language in the sample. Albania receives the least criticism overall, with the lowest hard-flag odds of any country (OR 0.75, $p < 0.001$). This pattern is visible at the paragraph level. A 2022 Serbia report paragraph on biomedicine notes that "laws on organ transplantation remain to be implemented since 2019" and that "administrative capacity remains very limited," yet contains no hard criticism terms whatsoever. The EU is clearly aware of the problem but stops short of condemnation. This is exactly what the low severity ratio captures: not an absence of concern, but a consistent choice to express that concern in softer terms (see Figure 3 in Appendix A).

4.3 Year trend

Despite the visible shift in enlargement logic after 2022, we find no systematic change in criticism severity across the sample period (M1 OR = 0.994, M2 OR = 1.021, M3 $\beta = 0.001$). Rather than a null result, we read this as meaningful. If geopolitical calibration had only entered EU language after the war, we would expect a post-2022 shift. The absence of one suggests the strategic differences between countries were already baked into how the EU wrote about each candidate long before 2022.

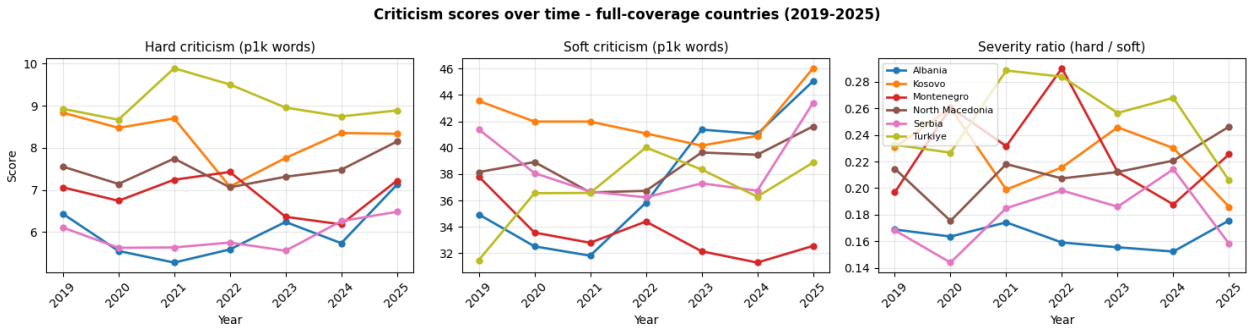

Figure 1: Hard criticism, soft criticism, and severity ratio across six EU candidate countries, 2019-2025

5 Conclusion

Our analysis reveals three main findings that speak directly to the theoretical debate about what drives the European Union's enlargement language. The first is about which topics attract the

harshest language. Even after controlling for country and year, judiciary and anti-corruption paragraphs are more likely to contain hard criticism than paragraphs on other topics, while economic paragraphs show no significant severity effect despite a modest positive association with hard criticism in absolute terms. This lines up with the Gancia et al. (2019) argument that deep integration is really about institutional convergence and not just market access. In practice, the EU is toughest exactly where it needs candidate countries to look most like itself: in their courts and their anti-corruption frameworks. While Dimitrova (2024) suggests that enlargement has increasingly become about geopolitics, our results show that the institutional bar has remained consistent, even after the geopolitical shift in 2022. The criteria do not change across countries, but how directly the EU reports on them does. The bar is uniform, but the language is not.

The second finding is that there is no time trend. Hard criticism has not systematically gone up or down across 2019–2025 once we controlled for which countries are in the sample and what topics their reports cover. The strategic differences between countries are already fully captured by the country fixed effects, not by the year trend. What the flat time trend tells us is that these country-level patterns were stable and consistent across the whole period, including before 2022. The EU was not suddenly more or less critical after the war; it was writing about each country in roughly the same way it always had. This is consistent with the argument in our introduction that geopolitical calibration was already operating quietly in EU language before the war made it explicit.

The third finding is that country differences survive controls for both topic and time. These gaps are not just a reflection of what problems each country has because controlling for topic composition would have explained them away. They reflect something about how the EU chooses to talk about each country. And because there is no time trend, these patterns were already in place before Russia's invasion changed the enlargement agenda. The strategic calibration that Economides et al. (2024) find in parliamentary enlargement discourse was present in the Commission's own progress reports well before the war made it explicit.

One thing we would really like to explore in future work, once enough annual reports have accumulated for the fast-tracked countries, is applying the same dictionary pipeline to Georgia and Ukraine and comparing their severity trajectories against the pre-2022 candidates, to test whether the linguistic patterns we detected here become even more pronounced when geopolitics openly drove the accession decision.

6 Limitations

Our main limitation is dictionary coverage. The keyword lists were built by hand and may miss ways of expressing criticism that do not appear in the initial list. Stemming mitigates this for inflectional variants, but semantically distinct phrasings that fall outside the dictionary contribute nothing to the score. If certain phrasings are more common in some countries' reports than others, this could introduce systematic bias into the cross-country comparisons rather than just adding random noise.

The second limitation is the use of the paragraph as the unit of analysis. Some assessments in the reports stretch across multiple paragraphs, so splitting at paragraph boundaries may fragment arguments that belong together. When this happens, the criticism score for each fragment will be lower than if the passage had been scored as a whole, since the relevant terms are spread across a

larger word count denominator. Our preprocessing robustness checks suggest this does not affect country rankings, but it likely compresses the absolute score levels symmetrically across countries.

Two smaller issues are also worth flagging. The method does not handle negation, so a phrase like "did not fail" would still count as a hit for *failed*. Standard negation handling using dependency parsing or a negation scope window would catch most of these cases, though EU institutional prose is formulaic enough that true negations of critical terms are rare. The method also does not distinguish between a term appearing in the Commission's own assessment and the same term appearing in a subordinate clause or in reported speech, a problem that token-level classification or named entity disambiguation could partially address. Both of these issues are more likely to affect overall score levels than the comparisons between countries that drive the main findings.

References

Enlargement and Eastern Neighbourhood Department of the EU. 2026. EU Enlargement Reports. Available at: https://enlargement.ec.europa.eu/enlargement-policy/strategy-and-reports_en.

Gancia, Gino, Giacomo A. M. Ponzetto, and Jaume Ventura. 2019. A Theory of Economic Unions. *Barcelona GSE Working Paper Series* No. 1110.

Dimitrova, Antoaneta. 2024. Dilemmas of EU Enlargement: Geopolitics, Conditionality, and Citizens' Concerns. Swedish Institute for European Policy Studies (SIEPS). Available at: https://www.sieps.se/....

Baker, Scott R., Nicholas Bloom, and Steven J. Davis. 2016. Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics* 131(4): 1593–1636.

Economides, Spyros, Kevin Featherstone, and Natasha Hunter. 2024. A Geopolitical Turning Point? Enlargement Discourse after the Russian Invasion of Ukraine. *JCMS: Journal of Common Market Studies*. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC12619659/.

A Additional plots

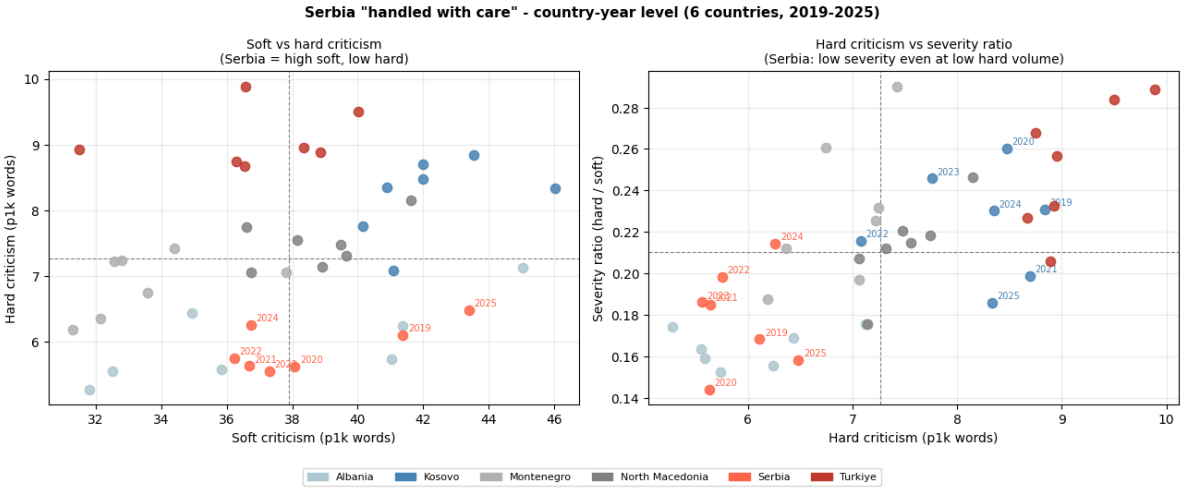

Figure 2: Country-year level scatter plots for the six full-coverage countries, 2019-2025. Left panel: Serbia clusters in the high-soft, low-hard quadrant across every year in the sample. Right panel: Serbia's severity ratio remains low even relative to countries with similar hard criticism volume, while Turkey sits in the top-right corner on both dimensions.