Is “*manovra*” really “*del popolo*”? Linguistic Insights into Twitter Reactions to the Annual Italian Budget Law[1]

Claudia Roberta Combei
University of Bologna
claudia.combei2@unibo.it

Abstract

English. Relying on linguistic cues obtained by means of structural topic modelling as well as descriptive lexical analyses, this study contributes to the general understanding of the Twitter users’ response to the annual Italian budget law approved at the end of December 2018. Some topics contained in the dataset of tweets are procedural or generic, but besides those, it often emerges that Twitter users expressed their concern with respect to the provisions of this law. Supportive attitudes seem to be less frequent. This paper also advocates that findings from inductive studies on Twitter data should be interpreted with caution, since the nature of tweets might not be adequate for drawing far-reaching generalisations.

1 Introduction

In the last decade, Internet has revolutionized human communication and interaction. And among all forms of digitally-mediated communication, social media stand out as one of the most effective. As Boulianne (2017) points out, the effects of social media depend on their nature of use (e.g. source of information; one-to-one/one-to-many/many-to-many communication; networking and relationship-building; expression of opinions; etc.).

Nowadays, potentially everyone with a computer or a mobile device having access to the internet can write and share contents which may be viewed and debated immediately by other people. The impact of a social media post may be huge, and unlike other prior forms of communication, it can easily cross borders in just a few seconds. In fact, social media make things happen faster than ever before. For instance, Facebook and Twitter were crucial in allowing the Arab uprisings or the Romanian anti-corruption protests to happen more efficiently and on a larger scale.

2 Tweets and politics

Besides their essential role in information dissemination, networking, and people mobilization, social media are also important indicators and predictors of their users’ opinions, sentiments and attitudes. In fact, various studies have explored people’s reactions towards social, economic, and political issues, by analysing social media posts (e.g. Burnap et al., 2014; Gaspar et al., 2016; Nesi et al., 2018), especially tweets, since they are easily retrievable by means of APIs.

With over 6,000 tweets posted every second, corresponding to roughly 350,000 per minute, 500 million per day, and around 200 billion per year, Twitter has become one of the main tools of communication worldwide (Internet Live Stats, 2019). The number of tweets written daily seems to be correlated to things happening in the real world, and, as a matter of fact, it was shown that important events generate high number of tweets (cf. Hughes and Palen, 2009), something that is generally reflected also on the Twitter “trends”. Based on Hootsuite’s (2019) report, each month, in Italy there are almost 2.5 million active users[2] of Twitter, a datum that confirms the popularity of this network among various layers of Italian audience.

[2] Users that write or share at least one tweet every month are defined “active”.

This means that Twitter may represent an easily exploitable opportunity for politicians in their attempt to reinforce communication with potential voters in what might be defined as a permanent digitally-mediated electoral campaign. Additionally, it has been suggested that Twitter could be used to model and predict public opinion and behaviour regarding political events, such as electoral campaigns (e.g. Coletto et al., 2015; Kalampokis et al., 2017). In fact, Ott (2017: 59) claims that Twitter may be the ideal tool for the afore-mentioned purposes since, it "*privileges discourse that is simple, impulsive, and uncivil*."

While indeed tweets have been widely used to analyse public opinion and political discussions in all its forms, several methodological considerations are dutiful. First of all, Twitter users do not represent an optimal sample for public opinion or voting population, especially due to their higher than average level of education and political sophistication, as well as a generally younger age (cf. Gayo-Avello, 2013; Barberá et al., 2015). As a matter of fact, we believe it is more accurate to define Twitter users as a potential share of electorate. Secondly, the language of tweets is characterised by succinctness and sometimes informality, colloquialism, irony, and susceptibility to rumour, all of which are aspects that render the results of large-scale analyses hard to interpret and generalise.

3 Aims and motivations

Acknowledging all the limitations mentioned above, this inductive exploratory study aims to contribute to the growing body of literature examining Twitter and its increasingly prominent role in online communication by studying its application in the context of political discourse. In particular, the linguistic approach presented here is providing insights into tweets regarding the discussion and the approval of the annual Italian budget law (in Italian "*legge finanziaria*" and/or "*legge di bilancio*"). This law was also often labelled as "the manoeuvre" (in Italian "*la manovra*") and "the people's manoeuvre" (in Italian "*la manovra del popolo*") by its proponents – in particular *Movimento 5 Stelle* (abbreviated *M5S*) –, mainly due to some of its populist provisions (e.g. the citizen's basic income and pension).

By means of structural topic modelling (cf. Roberts et al., 2014) and descriptive analyses (i.e. terminology extraction of multi-keywords and word sketches), we are interested in grasping the Twitter users' attitudes towards the budget law in a significant moment for the first populist Government in the eurozone, namely the coalition formed by *Lega* and *M5S*.

This topic is worth studying since the two parties displayed differences in economic, fiscal, infrastructural, and social policies both in the electoral campaign for the 2018 general elections as well as during the first months of government. For instance, *Lega* supported the flat taxation on incomes, while *M5S* the citizen's basic income ("*reddito di cittadinanza*" in Italian). However, these measures, although slightly modified, as well as the amendment to the 2011 pensions reform ("*quota 100*" in Italian) were included in the coalition agreement and subsequently in the draft for the annual budget law. The bill also contained various other economic and fiscal provisions (e.g. taxes on digital services; new VAT rates; reducing military expenses and the Italian contribution to United Nations; new labour measures; environmental incentives; etc.)[3].

We believe that the textual material contained in tweets may be promising in providing hints on how Twitter users – a fraction of the Italian voters – reacted to the provisions of the budget law. Linguistic insights into tweets might be able to guide us in understanding whether the so-called "*manovra del popolo*" was perceived by Twitter user as representing indeed the people's interest.

4 Data

Although in the Western world there are three mainstream social media networks (i.e. Facebook, Instagram, and Twitter), in this paper we analyse Twitter posts, primarily as a consequence of data availability. Indeed, unlike other tools for social media, Twitter APIs for R (R Core Team, 2018) allow scholars to collect large quantities of tweets and their related metadata in a rather effortless way.

Using the *rtweet* package (Kearney, 2019) for R and Twitter's developer account, we collected a dataset of 167,259 Twitter posts, for a total of 6.5 million tokens, consisting in tweets and retweets

[3] The full text of the annual Italian budget law (*Legge 30 dicembre 2018, n. 145 – Bilancio di previsione dello Stato per l'anno finanziario 2019 e bilancio pluriennale per il triennio 2019-2021*) was published on the Official Gazette of the Italian Republic (GU n.302 31-12-2018 - Suppl. Ordinario n. 62) and it is available online at this webpage: https://www.gazzettaufficiale.it/atto/stampa/serie_generale/originario (accessed on the 1st of June 2019).

related to the Italian budget law. Moreover, we extracted 88 metadata describing the tweet (i.e. character length, device used, number of retweets, etc.) and the user (i.e. username, location, gender, etc.). In order to capture the most important phases of the Twitter discussion about the annual budget law and considering the one-week rate limit for tweets extraction imposed by the Standard Search API[4], the data were collected weekly from the 27th of November 2018 through the 8th of January 2019, for a total of 43 consecutive days. The hashtags used as keywords in the queries represented all the names given to the budget bill by Italian political actors, the press, and the public opinion: "*#leggedibilancio*", "*#leggefinanziaria*", "*#manovra*", "*#manovradibilancio*", "*#manovraeconomica*", "*#manovradelpopolo*", and "*#manovrafinanziaria*". This guaranteed a large coverage of Twitter users and tweet typologies. Some of the afore-mentioned hashtags (e.g. "*#manovra*", "*#manovradelpopolo*") were also trending at the end of December.

To avoid duplicates, we discarded all retweets and all posts that contained quotes of other tweets. The removal process was obtained by filtering the dataset, thus selecting only tweets whose values for "is_retweet" and "is_quote" corresponded to "FALSE". Duplicates other than retweets and quotes were removed with R's base functions *duplicated* – which identified duplicated tweets – and *unique* – which extracted unique tweets. Since the aim of this study is to uncover the reactions of the Italian voters active on Twitter, we removed the tweets written by political actors. To do so, we defined a list containing the Twitter usernames of the members of the Italian Parliament, as well as those of the official national and local party profiles; this list was used to automatically filter and remove tweets published by the unwanted profiles. We decided to keep tweets from news agencies, online newspapers, and television channels, since they could represent vectors of information exchange regarding the topic analysed in this study. The final dataset contained 20,891 tweets.

Tokens	701,986
Words	414,803
Types	75,485
Lemmas	31,947

Table 1: Dataset statistics.

4.1 Pre-processing

Since the tweets and their metadata would have been used for lexical analyses and structural topic modelling[5], we performed several pre-processing steps: defining a "stop words" list for Italian consisting of roughly 1,000 lexically empty or uninformative words (i.e. prepositions, conjunctions, auxiliary verbs, etc.); uniformizing, normalising and cleaning the texts with various corpus processing functions available on the R packages *quanteda* (Benoit et al., 2018), *tm* (Feinerer, Hornik, and Meyer, 2008), and *qdapRegex* (Rinker, 2017). Hashtags at the beginning and inside the tweet sentences were kept and decomposed into words (i.e. from "*#trasportipubblici*" to "*trasporti pubblici*"), while those after the final point were removed, since most of the times they represented one of the keywords used for extracting tweets. Numbers, punctuation, sequences made up of a single character, and excessive white spaces were removed as well. In order to further use temporal metadata as a covariate for the topical prevalence, the "created_at" metadatum was divided it into date and hour.

5 Analyses and results

As a result of the ever-growing interest and availability of text data – often unstructured –, various statistical and machine-assisted approaches for the analysis of textual material have been proposed. In this paper we are employing the Structural Topic Model (STM) – a generative model of word counts – (cf. Roberts et al., 2014) in R to discover topics from tweets on the annual Italian budget bill and to estimate their relationship to temporal metadata.

Similarly to Latent Dirichlet Allocation (cf. Blei, Ng, and Jordan, 2003) and Correlated Topic Model (cf. Blei and Lafferty, 2007), in the STM approach, a topic represents a mixture over words where each word has a probability of pertaining to a topic, whilst a document is a mixture over topics, therefore a specific document can consist of various topics. The sum of the topic proportions across topics for a specific document as well as the sum of word probabilities for a given topic both qual to 1. The main innovation of STM is the possibility to model topical prevalence and topical content[6] as a function of metadata. Here we are

[4] A description of the Standard Search API for Twitter is available at this webpage: https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html (accessed on the 1st of June 2019).

[5] Considering the scope of this paper and the analyses proposed, emoticons and emojis were left out.

[6] The topical prevalence shows the frequency with which a specific topic is discussed, while the topical

using the date covariate to explain topical prevalence over time.

5.1 Topics

After having employed the STM's *searchK* function to perform several tests, such as held-out likelihood and residual analysis, the ideal number of topics seemed to be between 10 and 14. Additionally, STM gave the possibility to set the type of initialization, so here the spectral one was chosen, since previous studies had proven its stability and consistence (cf. Roberts, Stewart, and Tingley, 2016). All results presented in this paragraph are based on a *K* of 10. The date of the tweet was used as a prevalence covariate; as a word profile we opted for the highest probability. We did not use the stemming function on STM since it did not perform well on Italian.

Figure 1 in Appendix shows the topics related to the annual Italian budget law as they emerged from the analysis of tweets. Each topic was further classified into one category (i.e. EU & Confidence, Main Measures, Criticism & Concern, Government vs. Opposition, Procedures – Generic, Support). This classification was based on the correlations obtained from a hierarchical clustering representation performed with the *plot* function of the *stmCorrViz* package (Coppola et al., 2016), on the review of the most characterising words, and on the examination of the most exemplar documents, namely the tweets that had the highest proportion of words associated with the topic.

Although we do not claim to model public opinion from tweets, interestingly, the topics managed to echo various issues regarding the budget law. Judging by the expected topic proportions, one could order the most prevalent topics as follows: Topics 9, 8, and 3 (sum of topic proportions: 0.29) reflect disapproval and doubts towards the provisions of the budget law; Topics 1 and 7 (sum of topic proportions: 0.22) describe the difficult negotiation with the European Union (EU) and the threat of an infringement procedure; Topics 10 and 2 (sum of topic proportions: 0.19) depict the main measures contained in the budget bill; Topic 6 (topic proportion: 0.13) illustrates the support to the budget bill and to the Government; Topic 5 (topic proportion: 0.11) refers to the procedures regarding the discussion, the vote, and the approval of the budget law; and Topic 4 (topic proportion: 0.06) reveals the conflict between the Government and the oppositions on the provisions of the law.

After having calculated the estimated effects of the temporal covariate on topical prevalence, a plot displaying this variation was created. Figure 2 in Appendix shows how the afore-mentioned topics varied over the 43 days considered. Topics are ordered as a function of their expected proportions.

Firstly, there emerged that the variation was not particularly strong, except for some topics. For instance, Topic 9 had a peak at the end of December/the beginning of January, suggesting that Twitter users might have written tweets of concern soon after the approval of the annual Italian budget law. On the other hand, Topic 6, which contained mostly tweets of support towards the measures of the budget bill seemed to be prevalent primarily at the end of November and in mid-December. The procedural topic was generally prevalent at the end of December, a timeframe corresponding to the vote and approval of the law. The two topics summarising the negotiations with the EU, the confidence, and the possible infringement procedure were pervasive during the entire period considered, with some peaks in early- and mid-December. Topic 4 that regarded the disagreement between the Government and the opposition was constant over time, and so were the topics delineating the main measures of the law.

5.2 Descriptive lexical analyses

We were also interested in performing descriptive lexical analyses on tweets. First of all, with the terminology extraction tool on Sketch Engine (Kilgarriff et al., 2014) we obtained multi-keywords – able to convey more insights than single words on the issues examined – that appear more frequently in our dataset than in the reference corpus (i.e. Italian Web 2016 – itTenTen16, cf. Jakubíček et al., 2013, for TenTen corpora). If we exclude the hashtags used as keywords for tweets extraction, these are the 30 most representative syntagmas in our dataset:

Syntagma	Translation into English
reddito di cittadinanza	the citizen's basic income
procedura di infrazione	infringement procedure
clausole di salvaguardia	safeguard clauses

content represents the words used to discuss about that topic (cf. Roberts et al., 2014: 1068).

voto di fiducia	confidence vote
blocco assunzioni	hiring freeze
professioni sanitarie senza titolo	health professions without a degree
flat tax	flat tax
commissione bilancio	budget committee
gilet azzurri	blue vests
taglio pensioni	pension cuts
scatoletta di tonno	tuna can
governi precedenti	previous governments
pensioni minime	minimum pensions
scatola chiusa	black box
nuove tasse	new taxes
promesse elettorali	campaign promises
fasce deboli	vulnerable citizens
deficit strutturale	structural deficit
accordo tecnico	technical arrangement
braccio di ferro	trial of strength
appalti senza gara	no-bid contracts
assurdità totale	total nonsense
terrorismo mediatico	media terrorism
auto inquinanti	polluting cars
più tasse	more taxes
governo sovranista	sovereignist government
manovra contro il popolo	manoeuvre against the people
false promesse	false promises
IVA sui tartufi	VAT for truffles
popolo italiano	Italian people

Table 2: The most representative syntagmas in the dataset.

It is clear that various multi-word expressions referred to procedural aspects, such as those reflecting the vote and the approval of the budget law (e.g. “confidence vote”), while others were used to list its measures, especially fiscal and economic policies (e.g. “the citizen’s basic income”, “flat tax”, etc.). Nevertheless, various syntagmas seemed to express doubts with respect to the provisions of this law. In fact, often, the words chosen by many Twitter users to express their criticism were rather strong (e.g. “total nonsense”, “black box”, “sovereignist government”, etc.).

These concerns and rather negative reactions to the budget bill were reflected also in the word sketches (i.e. visual representations of collocations and word combinations obtained on Sketch Engine) for the words “*manovra*” and “*legge*”. Generally, three different scenarios are distinguishable.

First of all, there were several neutral verbs, nouns, and modifiers associated to the budget law, most of which regarding its procedural aspects. The most frequent (i.e. frequency ≥ 10.81 per million) are listed below:

Word/Syntagma	**Translation into English**
scrivere	write
cambiare	change
modificare	modify
discutere	discuss
approvare	approve
contenere	contain
prevedere	consist
varare	launch
votare	vote
passare	pass
riscrivere	rewrite
promulgare	promulgate
gialloverde	yellow-green
economica	economic
finanziaria	financial
populista	populist
discussione	discussion
commissione	commission
bilancio	budget

Table 3: Neutral associations.

Next, some positive evaluations of the budget law emerged. The most frequent (i.e. frequency ≥ 10.81 per million) are listed below:

Word/Syntagma	**Translation into English**
favorire (*l’innovazione*)	favour (innovation)
grande	big
buona	good
bella	beautiful
significativa	significant
del popolo	of the people
del cambiamento	of the change
per i cittadini	for the citizens
per la crescita	for the growth

Table 4: Positive associations.

Nonetheless, several word associations seemed to suggest negative reactions to the budget law. The most frequent (i.e. frequency ≥ 10.81 per million) are shown below:

Word/Syntagma	**Translation into English**
recessiva	recessive
piena di errori	full of errors

dannosa	dangerous
cattiva	bad
iniqua	unfair
scellerata	wicked
sbagliata	wrong
snaturata	wretched
taroccata	false
vuota	empty
assurda	absurd
folle	deranged
truffa	fraud
contro il popolo	against the people
del popolino	of the masses
del cappio	of the noose
da lacrime	tearful
scontro	dispute
protesta	protest
vergogna	shame
bocciatura	failure
della povertà	of the poverty
dell'assistenzi-alismo	of welfarism
buio	dark
diminuire	diminish
tagliare	cut

Table 5: Criticism associations.

Finally, using the *tm*'s *findAssocs* function, we calculated the associations of the lemma "*manovra*" in the term-document matrix; some of the afore-mentioned criticism words (e.g. "absurd", "recessive", "bad") had a correlation higher than 0.03, suggesting a rather frequent co-occurrence.

6 Conclusions

This paper explored the Twitter users' reactions to the annual Italian budget bill. STM outputs and descriptive lexical analyses showed that tweets concerned various aspects associated to the object of this study. Apart from talking about procedural and generic issues, users expressed their doubts and disapproval with respect to the measures of the budget law. Generally, tweets supporting this law were less frequent. The findings of this study, although preliminary, might be seen as indicators of what subsequently turned out to be a failure for the first Conte government. Still, as reiterated throughout the paper, the results might not reflect the real attitudes of the Italian voting population, since Twitter users tend to be younger and to have an above the average level of education and political sophistication (cf. Barberá et al., 2015). Moreover, tweets, by nature, might not be suitable for drawing steady generalizations, even if the prospects they offer for content and discourse analysis are indeed significant. Further research on this topic might include the investigation of Twitter user's reactions by means of sentiment analysis.

References

Pablo Barberá, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. Tweeting from Left to Right. *Psychological Science*, 26(10):1531–1542.

Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

David M. Blei and John D. Lafferty. 2007. A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1):17–35.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(3):993–1022.

Shelley Boulianne. 2017. Revolution in the making? Social media effects across the globe. *Information, Communication & Society*, 22(1):39–54.

Pete Burnap, Matthew L. Williams, Luke Sloan, Omer Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, and Alex Voss. 2014. Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining*, 4(206):1-14.

Mauro Coletto, Claudio Lucchese, Salvatore Orlando, and Raffaele Perego. 2015. Italian Information Retrieval Workshop - IIR 2015. In *Proceedings of the 6th Italian Information Retrieval Workshop*, Cagliari. CEUR Workshop Proceedings.

Antonio Coppola, Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. 2016. *stmCorrViz: A Tool for Structural Topic Model Visualizations*. R package version 1.3. Retrieved from https://cran.r-project.org/web/packages/stmCorrViz/index.html/ (accessed on the 1st of June 2019).

Ingo Feinerer, Kurt Hornik, and David Meyer. 2008. Text Mining Infrastructure in R. *Journal of Statistical Software*, 25(5):1-54.

Rui Gaspar, Cláudia Pedro, Panos Panagiotopoulos, and Beate Seibt. 2016. Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events. *Computers in Human Behavior*, 56:179–191.

Daniel Gayo-Avello. 2013. A Meta-Analysis of State-of-the-Art Electoral Prediction from Twitter

Data. *Social Science Computer Review*, 31(6):649–679.

Hootsuite Media Inc. 2019. Digital in 2019. Retrieved from https://hootsuite.com/it/risorse/digital-in-2019-italy (accessed on the 1st of June 2019).

Internet Live Stats. 2019. Twitter Usage Statistics. Retrieved from https://www.internetlivestats.com/twitter-statistics/ (accessed on the 1st of June 2019).

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychly, and Vít Suchomel. 2013. The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL 2013*. Lancaster, 125-127.

Evangelos Kalampokis, Areti Karamanou, Efthimios Tambouris, and Konstantinos Tarabanis. 2017. On Predicting Election Results using Twitter and Linked Open Data: The Case of the UK 2010 Election. *Journal of Universal Computer Science*, 23(3):280–303.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

Michael W. Kearney. 2019. *rtweet: Collecting Twitter Data.* R package version 0.6.9 Retrieved from https://cran.r-project.org/package=rtweet (accessed on the 1st of June 2019).

Amanda Lee Hughes and Leysia Palen. 2009. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3/4):248–260.

Paolo Nesi, Gianni Pantaleo, Irene Paoli, and Imad Zaza. 2018. Assessing the reTweet proneness of tweets: predictive models for retweeting. *Multimedia Tools and Applications*, 77(20):26371–26396.

Brian L. Ott. 2017. The age of Twitter: Donald J. Trump and the politics of debasement. *Critical Studies in Media Communication*, 34(1):59–68.

R Core Team. 2018. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/ (accessed on the 1st of June 2019).

Tyler W. Rinker. 2017. *qdapRegex: Regular Expression Removal, Extraction, and Replacement Tools.* R package version 0.7.2. University at Buffalo. Buffalo, New York. Retrieved from http://github.com/trinker/qdapRegex/ (accessed on the 1st of June 2019).

Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4):1064–1082.

Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. 2016. Navigating the Local Modes of Big Data: The Case of Topic Models. In R. Michael Alvarez (editor), *Computational Social Science: Discovery and Prediction (Analytical Methods for Social Research)*, 51–97. Cambridge University Press., Cambridge.

Appendix

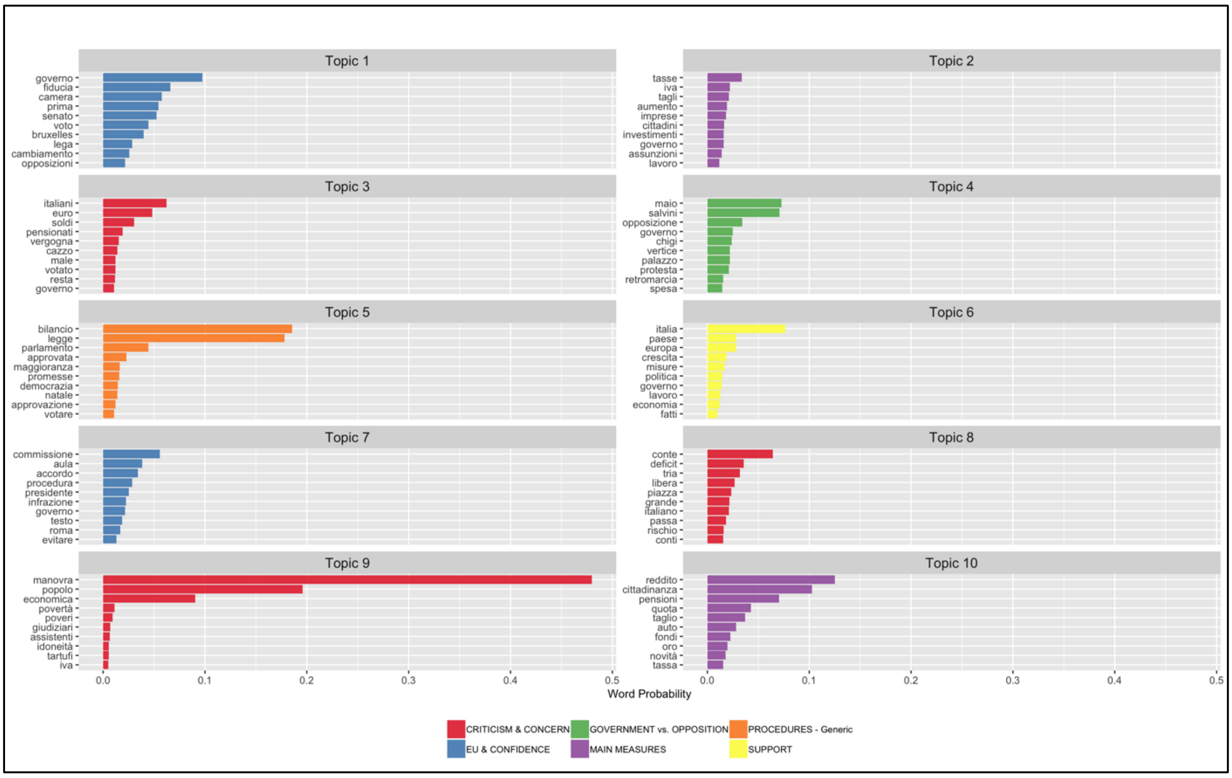

Figure 1: Topics and word probabilities.

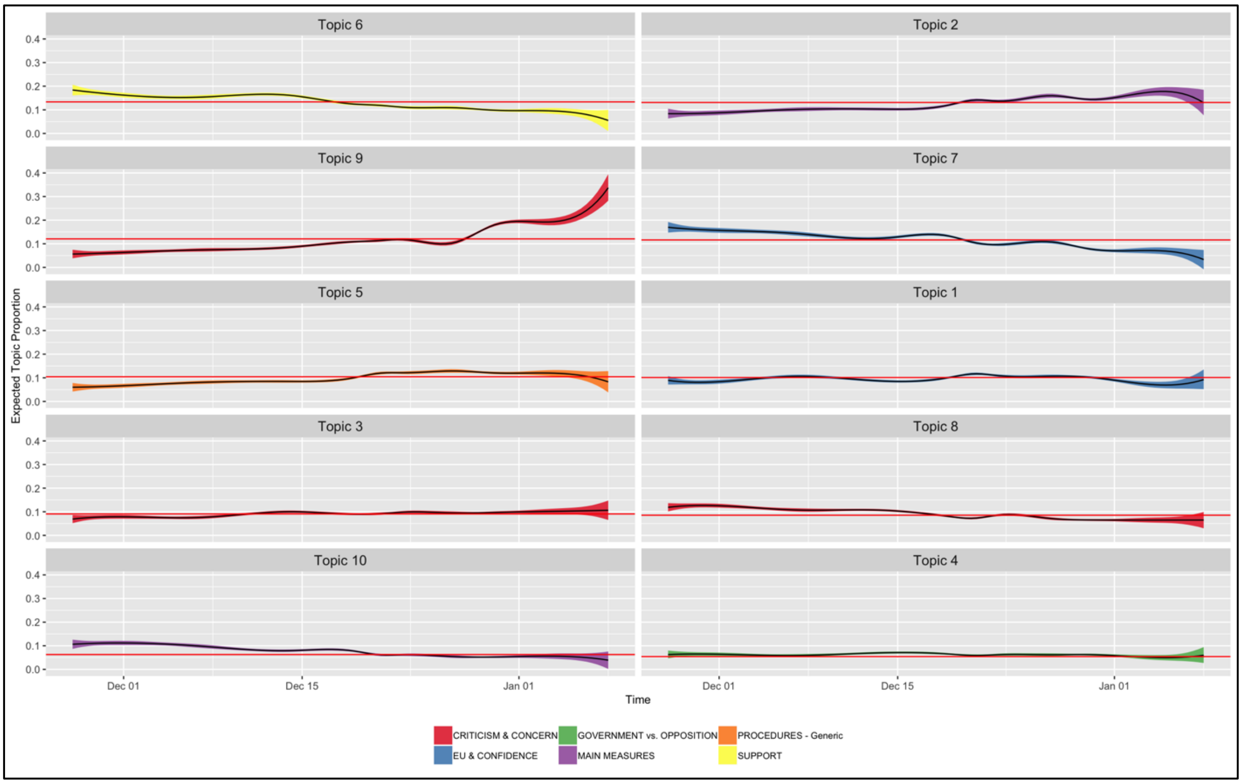

Figure 2: Variation of topic proportions over time.