What does Twitter have to say about ideology?

Sarah Djemili, Julien Longhi, Claudia Marinica, Dimitris Kotzinos, Georges-Elia Sarfati

To cite this version:

Sarah Djemili, Julien Longhi, Claudia Marinica, Dimitris Kotzinos, Georges-Elia Sarfati. What does Twitter have to say about ideology?. NLP 4 CMC: Natural Language Processing for Computer-Mediated Communication / Social Media - Pre-conference workshop at Konvens 2014, Oct 2014, Hildesheim, Germany. pp.16-25. halshs-01058867v1

HAL Id: halshs-01058867

https://shs.hal.science/halshs-01058867v1

Submitted on 28 Aug 2014 (v1), last revised 8 Oct 2014 (v2)

What does Twitter have to say about ideology?

Sarra Djemili
ETIS - UCP/ENSEA/CNRS 8051
UCP, Cergy-Pontoise, France
sarahdjemili@yahoo.fr

Julien Longhi
CRTF - UCP/EA 1392
UCP, Cergy-Pontoise, France
Julien.Longhi@u-cergy.fr

Claudia Marinica
ETIS - UCP/ENSEA/CNRS 8051
UCP, Cergy-Pontoise, France
Claudia.Marinica@u-cergy.fr

Dimitris Kotzinos
ETIS - UCP/ENSEA/CNRS 8051
UCP, Cergy-Pontoise, France
Dimitrios.Kotzinos@u-cergy.fr

Georges-Elia Sarfati
STIH/ EA 4509
Paris Sorbonne, France
georgesarfati@gmail.com

Abstract

Political debates bearing ideological references exist for long in our society; the last few years though the explosion of the use of the internet and the social media as communication means have boosted the production of ideological texts to unprecedented levels. This creates the need for automated processing of the text if we are interested in understanding the ideological references it contains. In this work, we propose a set of linguistic rules based on certain criteria that identify a text as bearing ideology. We codify and implement these rules as part of a Natural Language Processing System that we also present. We evaluate the system by using it to identify if ideology exists in tweets published by French politician and discuss its performance.

1 Introduction

Political and ideological debates have been a part of our political and societal functions for many years, to some extend since the first steps of the civilization. One could argue that the opinions of others are important to us in order to make for example a responsible decision regarding the electability of a particular candidate, to look beyond appearances and be able to judge the character of people. This includes evaluating their intelligence and leadership abilities, but it also involves learning about people's stance on various issues. On the other hand, fewer people have anymore the time and want to put the effort to go through the analysis of short or longer texts that position people and opinions or even worse sometime even reading them does not provide adequate answers. Moreover, the explosion of the internet brought multiple ways of communicating one's political opinions, thus making the whole process more difficult. In this context, microblogging services like the Twitter network give people the ability to express themselves with brevity but with speed and with less preparation thus exposing them more easily into the public. So, identifying or even studying ideology has become an even more challenging task (Riabinin, 2009).

Apart from that, studying ideology has always been a main issue in French discourse analysis domain. However, a semantic analysis of ideology has not been fully and rigorously developed (see Rastier 's assessment in (Rastier, 2011)), so even nowadays, these analyses lack of scientific description and especially rigorous evaluation. In that respect, one of the objectives of this article is to provide rigorous criteria for the identification of ideologies in tweets but also to implement them in a tool which allows their identification and validation. The complementarity with research in computer science provides answers to longstanding questions in the literature of discourse analysis. The choice of working on Twitter is justified by the fact that it is characterized as a new genre of political discourse as we showed in (Longhi, 2013), and due to its brevity it reflects a semantic condensation possibly to be favorable to ideologies. The work presented here is evaluated over text (tweets) that are in French, which was an obvious choice given the fact that the authors live and work in France and that we draw the rules we propose from criteria suggested for text in French. Apparently similar approaches could exist in other languages; transferring though either the criteria or the rules or both does not seem to work given the particularities in each language and the fact that our work is based on expressing and quantifying linguistic rules.

Political discourses were already analyzed in the literature, but this area is still young especially when the object of research is text produced in social media environments and when additionally we aim to identify relevant tweets based on the existence of ideological references in them. Some existing studies focus on discovering political affiliations in informal web-based contents like news articles (Zhou et al., 2011), political speeches (Dahllf, 2012) and web documents (Durant and Smith, 2007; Durant and Smith, 2006; Efron, 2006). Political datasets such as debates and tweets are explored for classifying users' positions (Walker et al., 2012; Somasundaran and Wiebe, 2010) and also for predicting election results (O'Connor et al., 2010) or the political party affiliation (Conover et al., 2011). These works use for prediction the content and other corpus specific properties such as hashtags, social networks, etc. Other works use ideological political beliefs for party prediction (Gottipati et al., 2013) exploiting likewise specific text properties.

Concerning ideology, a recent work, in (Iyyer et al., 2014), introduces a model for political ideology detection using a recursive neural network (RNN) in order to detect ideological influence at sentence level. The authors state that the resulting model can correctly identify ideological influence in complex syntactic constructions. However, they developed their own political ideology dataset annotated at phrase level for their model, and using it on another support such as social media (for example Twitter) may not be as effective.

In this work, we propose a set of rules that can be used to identify ideology in tweets and other short text messages. These rules stem from Sarfati's work (2014) on the necessary criteria to classify text as bearing any kind of ideology. On top of that we implemented these rules as part of a Natural Language Processing System that allows its use over the large corpuses that can be collected e.g. from Twitter. We evaluated these rules using actual tweets from French politicians.

This paper is structured as follows: in the next section we present Sarfati's criteria and we describe the steps taken to transform them to linguistic rules. The we describe how we implement these rules as part of a Natural Language Processing (NLP) System which we detail more in the beginning of the section (section 3). In section 4 we evaluate the implemented rules over a carefully validated corpus of tweets and present our preliminary results and first conclusions. We conclude the paper in section 5 by providing a sum up of the work so far and some pointers for future research.

2 From Sarfati's criteria to linguistic rules

The main objective of this paper is to detect whether or not a tweet is an ideology tweet, but not to classify it further according to the ideological references it carries. The work introduced by Sarfati (2014) provides the definition of the necessary criteria for a text to be classified positively as an ideology bearing text. Our effort is to transform the proposed criteria into linguistic rules and implement them as part of a Natural Language Processing System. Sarfati describes seven criteria on ideology: some of them are used just to characterize the type of the ideology or to describe it generally, but others are more definitive, permitting to detect ideology in text. Thus, in this study we concentrate on the five criteria presented below:

- Criterion 1: the deictic scope of the ideology is the one of a discourse state pretending to erase any clutch mechanism, any dependence on an enunciation place or any spatiotemporal context. The ideological discursive state claims *timelessness*;

- Criterion 2: the level of heterogeneity of the ideology consists in the negation itself of the mixed discourse, since under its strategic claim of transparency (universality) and of timelessness (transhistorical), ideology is structured as a *homogeneous* discourse, discursively smooth;

- Criterion 3: the ideology aims to produce the illusion of *timelessness* and it states an effective relevance for all times;

- Criterion 4: the reflexiveness level of the ideology consists in the fact of not pretending referring only to itself, that is to say that the ideology is its own end;

- Criterion 5: the ideology is *polychronous* as it pretends grouping all the temporal perspectives and canceling them.

More precisely, below we will describe how we transformed these criteria in rules in order to be use in our system. This transformation falls within the framework of the theory of discursive objects, developed by Longhi in (2008) for the concept of discursive object and in (2014) for the theory itself: one goal of this theory is to assign formal markers to discursive operations, in order to provide discourse analysis from pragmatic and enunciative criteria. More generally, this theory opens the theory developed by Sarfati to corpus linguistics.

Criterion 1:

Rule 1: no spatio-temporal deixis marks, such as: here (*ici* - fr), there (*là-bas* - fr), now (*maintenant* - fr), tomorrow (*demain* - fr), etc.

Rule 2: no interlocution subjects, such as: I (*je* - fr), you (*tu*, *vous* - fr), we (*nous* - fr), and occurrence of non-subjects, such as: he/she (*il/elle* - fr).

Rule 3: no proper nouns specifying places, people or factual data that are too precise.

Criterion 2:

Rule 4: in order to validate the universality and the homogeneity characteristics, no modalization marks should occur, such as: to seem to (*sembler* - fr), to appear (*paraître* -), to be able to (*pouvoir* - fr), to have to (*devoir* - fr). These marks outline speaker's attitude towards the statement. Moreover, this rule is confirmed also by the absence of punctuation marks such as "?" and "!" outside of a reported speech.

Rule 5: reduce the argumentation: no argumentative connectors, such as: but (*mais* - fr), so (*donc* - fr), because (*parce que*, *puisque* - fr), etc.), or neutral connectors, such as: and (*et* - fr), moreover (*de plus* - fr), etc.

Criterion 3:

Rule 6: for timelessness, the verb should be at present tense stating out a general truth. The past and future tenses should be present less frequently.

Criterion 4:

Rule 7: referring only to itself, the ideology should not contain other discourse marks, such as: double quotes, according to (*selon* - fr), as X says/thinks (*comme X dit/pense* - fr), etc.

Criterion 5:

Rule 6 seems adequate in order to validate this criterion.

3 Integrating linguistic rules in Natural Language Processing tools

The rules described in the previous section should help us to detect if a tweet is ideological or not. In order to develop a system implementing these rules, we evaluate the possibility of pushing the linguistic rules into the existing computer tools of Natural Language Processing (NLP).

Moreover, the implementation of the previous rules in our system requires a morpho-syntactic analysis in order to determine the part-of-speech category for each word in the tweet: a verb, an adjective, a noun, a preposition, etc. For this we also need to use a suite of NLP tools that carry the corresponding functionality.

To this end, we made a state of the art of the available open source (we surveyed only open source APIs both because it is open to anyone to use but also because the code is available for us to extend as needed) NLP APIs that we will detail next.

3.1 Morpho-syntactic analysis in NLPs

Part-of-speech (POS) tagging is one of the most fundamental parts of the linguistic analysis, a basic form of syntactic analysis which has important applications in NLP. The goal of this study is to analyze the POS tagging APIs available for french language and to compare them in order to see capabilities and limits for each one and to finally choose one or more to use. In our study, we are searching for the following elements: verb tenses, adjectives and nouns objective/subjective, personal pronouns, connectors, proper nouns, tense and time markers. We tested and evaluated three well-known POS taggers:

- Stanford POS Tagger[1]: this is a java implementation of the log-linear POS tagger which belongs to the Stanford NLP group. The provided library allows the user to tag the words in the text. The tagger has to load a trained file (named model) containing the necessary information for the tagger. There are several trained models provided by Stanford NLP group for different languages, including French; for French, the model is based on the pre-labeled French corpus named Treebank.

- Apache Open NLP[2]: the Apache Open NLP library is a machine learning based toolkit for the processing of natural language text. It supports the most common NLP tasks, such as tokenization, sentence segmentation, POS tagging, chunking, etc. These tasks are usually required to build more advanced text processing services. The french model is also based on Treebank corpus.

- Wikimeta[3]: is a labeling tool based on NLGbAse content. NLGbAse is a system producing metadata and components for natural language processing, semantic analysis, and labeling tasks. NLGbAse transforms encyclopedic text contents into structured knowledge fully integrated with the Linked Data network and the Semantic Web. NLGbAse metadata are used to produce resources and corpus training for information extraction tools like Wikimeta. Wikimeta detects named entities, and links them to their RDF description in the Linked Data Network. The semantic labeling web service API provides a REST-compliant, unique access point for all text-mining and content analysis functionality. The French java API of Wikimeta provides also a POS Tagger, named TreeTagger, and a frequency analysis tool.

We decided to test both Stanford POS tagger and Wikimeta java APIs (we didn't continue with Apach Open NLP as it is using the first tagged French corpus - Treebank) and compare their results.

Given the same text, we concluded that Stanford POS tagger was showing inferior performance compared to Wikimet's POS tagger. Indeed, the former takes in charge a reduced number of tags available on TreeBank, while the latter uses a wider list of tags (about 37). More precisely, during our evaluation, Stanford POS tagger didn't tag accordingly verbs' tenses, articles and amounts, that Wikimeta tagged correctly. The Table 1 presents a comparison between the 2 APIs applied on the following 2 tweets:

Tweet 1: *234 personnes au Raincy pour débattre du projet de l'UMP avec Bruno LEMAIRE, délégué général de l'UMP en charge du projet pour 2012.*

Tweet 2: *Débat primaires : ils font faux, remplis de haine contenue ! S'il n'y avait pas la caméra, il y aurait beaucoup d'éclats !*

Wikimeta allowed us to detect most of the elements that we need to implement our linguistic rules, such as: verbs tenses, connectors, proper nouns, personal pronouns. Based on the results of such kind of experiments we chose to use Wikimeta's API to develop our system for detecting ideological tweets.

3.2 Integration of rules

In this section, we detail how we integrate in our Wikimeta based system the linguistic rules that we created starting from Sarfati's criteria in section 2 and which technical issues this development introduces.

Rule 1: In order to implement this rule, we use initially Wikimeta to analyze the tweet as it provides three interesting tags: NTIME, NDAY and NMON which detect temporal entities. Then, given that we are interested in seventeen (17) spatio-temporal markers, we make a list of words containing all these markers and check if the tweet text contains it. For example, now (*maintenant* - fr), tomorrow (*demain* - fr), etc.

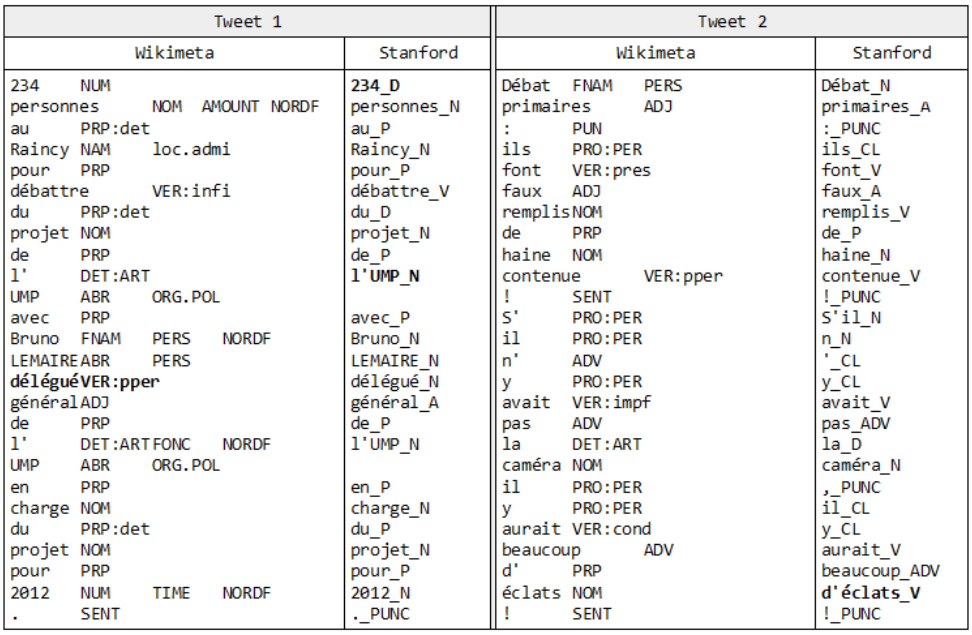
Rule 2: Similarly, for interlocution subjects, using Wikimeta we can easily check if the tweet's text contains: I (*je* -fr), you (*tu*, *vous* -fr), we (*nous* -fr), me (*moi* - fr), etc.

Rule 3: For this rule, Wikimeta can spot all proper nouns existing in the tweet, thus we just

[1] http://nlp.stanford.edu/software/tagger.shtml

[2] https://opennlp.apache.org/

[3] http://www.wikimeta.fr/



Tweet 1		Tweet 2	
Wikimeta	Stanford	Wikimeta	Stanford
234 NUM	**234_D**	Débat FNAM PERS	Débat_N
personnes NOM AMOUNT NORDF	personnes_N	primaires ADJ	primaires_A
au PRP:det	au_P	: PUN	:_PUNC
Raincy NAM loc.admi	Raincy_N	ils PRO:PER	ils_CL
pour PRP	pour_P	font VER:pres	font_V
débattre VER:infi	débattre_V	faux ADJ	faux_A
du PRP:det	du_D	remplisNOM	remplis_V
projet NOM	projet_N	de PRP	de_P
de PRP	de_P	haine NOM	haine_N
l' DET:ART	**l'UMP_N**	contenue VER:pper	contenue_V
UMP ABR ORG.POL		! SENT	!_PUNC
avec PRP	avec_P	S' PRO:PER	S'il_N
Bruno FNAM PERS NORDF	Bruno_N	il PRO:PER	n_N
LEMAIREABR PERS	LEMAIRE_N	n' ADV	'_CL
déléguéVER:pper	délégué_N	y PRO:PER	y_CL
généralADJ	général_A	avait VER:impf	avait_V
de PRP	de_P	pas ADV	pas_ADV
l' DET:ARTFONC NORDF	l'UMP_N	la DET:ART	la_D
UMP ABR ORG.POL		caméra NOM	caméra_N
en PRP	en_P	il PRO:PER	,_PUNC
charge NOM	charge_N	y PRO:PER	il_CL
du PRP:det	du_P	aurait VER:cond	y_CL
projet NOM	projet_N	beaucoup ADV	aurait_V
pour PRP	pour_P	d' PRP	beaucoup_ADV
2012 NUM TIME NORDF	2012_N	éclats NOM	**d'éclats_V**
. SENT	._PUNC	! SENT	!_PUNC

Table 1: Comparison between the results provided by Stanford POS tagger and Wikimeta.

search for words tagged with ”NAM” tag. Since proper nouns can be represented by abbreviations, Wikimeta can also help since it detects abbreviations and labels them with the ”ABR” tag.

Rule 4: To check if a tweet contains one of the four modal verbes, we first need to convert conjugated verbs to an infinitive form. To do that, we use a second API[4] developed by the Natural Language Processing group of Sheffield University, which ensures the lemmatization. Thus, we can compare the returned verb with the four (4) ones in our list. Concerning the question (?) and exclamation (!) marks, we just check if the tweet contains them or not.

Rule 5: Concerning the use of connectors, we look for the argumentative ones referring to a pre-existing list.

Rule 6: For rule 6, we use Wikimeta in order to detect the tense of each verb in the tweet. But, since a text can contain at the same time verbs at different tenses, we have to compute the predominant verb tense in the tweet. In order to do that, we count the apparition of each verb tense in the tweet by using three classes corresponding to past, present and future tenses.

Rule 7: Detecting discourse markers in French language was addressed by several works such as (Poulard et al., 2008; Giguet and Lucas, 2001; Buvet, 2012; Mourad and Desclés, 2003). The automatic identification of citations is not an obvious task as the identification of marks of reported speech, especially in the indirect case, is based on combinatorial heterogeneous linguistic units (Buvet, 2012). Authors proposed in (Giguet and Lucas, 2001) a syntactic strategy that we exploit. It consists of locating, without the need for exhaustive lists of shapes, three unknown elements: the source (of the citation - speaker), the reported speech and the text introducing the reported speech (e.g.: declared that (*a déclaré* -fr)). They used phrase-oriented criteria as computing indices: typographical signs (punctuation, capitalization), and morpho-syntactic and position-based elements for computing a three-value variable: source, reported speech and the introductory text. For that, they established a model for French corpus admitting two designs, according to the two different types of speech - direct or indirect - detailed in the following:

- the first one is a direct speech with the form *X explained that...* (X a expliqué que...);
- the second one is a indirect speech with the form ...*explained X* (...a expliqué X).

[4] http://staffwww.dcs.shef.ac.uk/people /A.Aker/activityNLPProjects.html

Moreover, for the direct speech, the double quotation mark outlines the opening of reported speech and the end of a reported speech (words in double quotes ” ”). For the indirect speech, he (*il* - fr) points out the presence of a speaker and that (*que* - fr) marks that a indirect reported speech might follow.

In tweets’ context, detecting direct speech is equivalent to identifying mentions having reply type (tweets that started with a @username) in addition to double quote signs. We also check the verbal speaker expressions. For indirect speech, markers like the ones mentioned above are identified. Additionally, we used the table given in (Mourad and Desclés, 2003) containing statistics about the most used verbs for detecting the speaker.

3.3 System operation

In order to apply the previous linguistic rules on an important number of tweets, we developed the system presented in Figure 1.

The system takes as input a set of political tweets and provides as result the rules that are satisfied by each tweet. To this end, a morpho-syntactic analysis is done on the tweets by Wikimeta API allowing POS annotation and detection of named entities.

4 Application to Twitter Dataset

4.1 Tweets

In recent years, social media activity has reached unprecedented levels. Hundreds of millions of users now participate in online social networks and forums, subscribe to microblogging services or maintain web diaries (blogs). Twitter, in particular, is currently the major microblogging service, with more than 255 million monthly active users who send more than 500 million Tweets (short text messages of up to 140 characters) per day[5]. They use tweets to report their current thoughts and actions, comment on breaking news and even engage in discussions.

4.2 Corpus Description

The corpus of tweets that we used was established by (Longhi et al., 2014) to serve two research projects: the ”CoMeRe” project which aims to establish a set of corpus-mediated communications networks, and the ”Digital Humanities and Data Journalism” project which aims to develop interdisciplinary research collaborations allowing to analyze political corpus produced via new ways of communication. The corpus was built starting from seven (7) French politicians of six (6) political parties. In order to generate political tweets, we started from a set of lists citing these politicians (7087 lists), and we selected those lists that have tweeted at least 6 times and which description contains the word *politics* - 120 lists remaining. Finally, 2934 tweets were recovered.

In order to be sure that we select politicians’ tweets (and not for example ones from journalists), we worked by keeping only the accounts cited in more than 12 lists; we have finally 205 politicians who were tweeting. For these 205 accounts we got the last 200 tweets of each on 27 March 2014 (34,273 tweets). This allows us to have a corpus focusing on the period between the two rounds of the 2014 municipal elections in France. For the less active accounts we took into account even earlier tweets because we wanted to keep the density of tweets from each account and the publication rate is not the same for all; the oldest tweet was sent on 2009-03-04 11:59:49).

4.3 Applying the rules

In this section we give some examples from the corpus of tweets to describe how our system processes tweets while applying the rules.

Tweet 1: *La loi DALO crée en 2007 un droit effectif au logement. Il faut pousser cette logique plus loin dans un service public du logement.*

Tweet 2: *Je suis ravi de pouvoir compter sur tous ceux qui m’ont accompagné ce soir sur Twitter pendant #motcroises, merci à vous !*

Tweet 3: *Bruno Lemaire : ”Les socialistes vivent dans le monde d’avant, c’est pourquoi nous devons inventer le monde d’après.”*

Tweet 4: *Le rassemblement ce n’est pas avoir peur les uns des autres, c’est être forts ensemble.*

In Tweet 1, Wikimeta allows to detect the temporal marker ”en 2007” and assign the tag NTIME to it. As a result, Tweet 1 does not satisfy Rule 1.

Tweet 2 does not satisfy Rule 1 neither because

[5] https://about.twitter.com/company

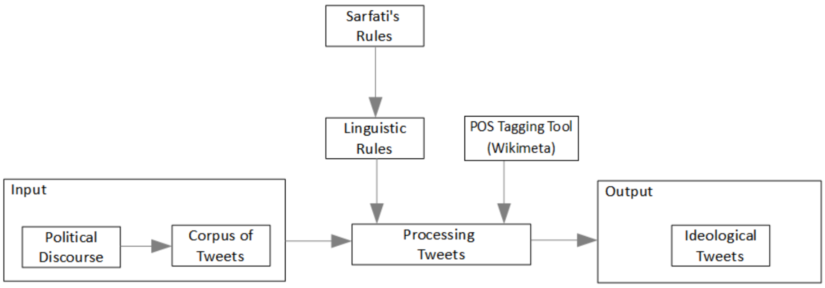


Figure 1: Ideological tweet detection system.

it contains the key word ”ce soir” (tonight - en). Tweet 2 contains also the proper noun ”Twitter” that Wikimeta is able to detect, so the Rule 3 is not satisfied. The tweet begins with an interlocution subject ”je” (I - en), that breaks Rule2. On other hand, the tense of verbs in the two tweets is the present, so Rule 6 is satisfied. They do not contain any connector which satisfies Rule 5. Tweet 2 contains an exclamation mark which is contradictory to Rule 4.

In Tweet 3, we have the modal verb ”devons” (must - en). The lemmatization gives the infinitive form of the verb so we can check that the Rule 4 is not satisfied in this case. Then, the tweet represent a direct speech where the relator is ”Bruno Lemaire” and the speech is between quotes and placed before quotation marks. So Rule 7 is not satisfied.

Tweet 4 did not contain any spatio-temporal mark or proper noun. It does not have an interlocution subject or any connector. The tense of verbs is the present and no exclamation or interrogation marks are seen. No modal verb is detected and no form of discourse exists. Tweet 4 satisfies the 7 rules and is identified as an ideological tweet by the system.

4.4 Results

For these preliminary results, we tested our system on 1188 tweets from the corpus, the system found that 14 tweets among them satisfy all 7 rules.

We analyzed these 14 tweets and noted the following observations:

- 4 tweets are too short to be effectively treated by the system (around 40 characters without the hyperlinks) and all of them indicate a hyperlink (e.g. *Quelques infos concernant les retraites http://t.co/zh2MHzct*);

- 3 tweets were not correctly classified by the system. This is due to misclassifying the verbs’ tense: in these 3 tweets we have the imperative tense tagged as present by Wikimeta (e.g.: *Arrêtons d’instrumentaliser les gay sur le dossier du mariage, demandons l’avis aux gens, proposons un référendum*).

- 1 tweet is not ideological, but the misclassification comes from the fact that a personal noun was omitted in the tweet. Indeed, the pronoun could be ”je” (I - en) or ”Il” (he - en) which is hard to automatically detect (*remercie les 500 personnes qui ont fait le déplacement pour l’inauguration de la permanence de campagne*).

- 6 tweets are validated as being ideological.

The detailed examination of the results shows a set of tweets that meets the expectations: actually, the ideology, as a state constructed by the discourse, is the product of a strategic decision, and the ideological reference is obviously the result of a rhetoric calculation.

In addition to the various criteria identified, it is important to note that these tweets contain a style that fits into a rhetorical and strongly argumentative reference in order to give a force to the tweet and to impose the ideology.

In this context, some structures are clearly identified:

Il faut (Have to - en):

* *Ce qu'il faut c'est établir des priorités, choisir des filières d'excellence, créer des emplois dans des secteurs porteurs.*

* *Il faut allouer plus de moyens éducatifs aux banlieues, en incitant par exemple statutairement et financiérement les enseignants en ce sens.*

* *Il faut permettre à ceux qui le souhaitent de prendre leurs responsabilités et de mener à bien des projets, dans leurs quartiers.*

Il y a (There is - en):

* *Il y a un problème de méthode pour règler les problèmes que rencontrent nos banlieues; il faut développer des conseils de quartier élus.*

A strong syntactic structure: topicalization such as *X, c'est x* or *ce qui est c'est* (X... is x...; which is...that is... - en):

* *Promouvoir un dialogue entre les équipes éducatives, les élèves et les parents, c'est ouvrir la voie à une école bienveillante.*

* *Ce qui est attendu des candidats ce ne sont pas des promesses, c'est un discours de vérité sur l'effort à produire #francebleu107_1*

The current hypothesis of detecting ideological tweets using semantic or/and syntactic elements can be enriched with style-based criteria, which could give interesting results; this point will be studied later in our research.

The evaluation of tweets, which don't satisfy several rules, is also interesting:

For example when applying Rule 4 on modalization, it might be interesting to evaluate the tweets containing the *devoir* (have to - en) verb, which in some cases do not necessarily indicates the involvement of the speaker, but rather a form of general truth, for example:

* *Les démocrates doivent s'unir pour mettre fin à cette violence dans le débat public. #BFMTV*

* *Un pays qui veut la réussite éducative de toutes et de tous doit valoriser ses professeurs. #Pisa2012*

Finally, more interesting for the rest of our work would be to discriminate different types of ideologies. For example, those who do not satisfy the rule 3 may correspond to a nationalist ideology, such as:

* *Quoi de plus naturel que l'amour de sa patrie ? Le patriotisme n'est pas un gros mot" #Souvenirfrançais*

But, as a general remark, the evaluation of the results indicates in the first place a good efficiency of applying linguistic rules for the detection of ideological tweets.

5 Conclusions and Future Work

In this paper, we proposed the implementation of Sarfati's criteria for detecting ideology on a text as a set of linguistic rules. Moreover, we provide a system that implements these rules as an extension of an NLP System. We use this system to experiment rules' implementation and their accuracy by analyzing a set of tweets of French politicians. The preliminary evaluation of the rules and their implementation give us encouraging results for the system's accuracy since most tweets identified as ideological were indeed so.

For the future we plan to extend the experimentation by using more tweets, to revise or relax some of the rules that might misclassify ideological tweets and to provide these rules as a standard extension to NLP systems so that they can be integrated in the everyday analysis of ideological discussions on social media.

6 Acknowledgements

This work is part of the Digital Humanities and Data Journalisme transdisciplinary project (funded by the Foundation of the Cergy-Pontoise University, France[6]) and of the CoMeRe project from group "Nouvelles formes de communication" of the consortium Corpus-écrits and supported by Corpus-écrits and Ortolang (Chanier et al., 2014).

References

Pierre-André Buvet. 2012. Traitement automatique du discours rapporté. In *Actes du colloque JADT 2012.*

Thierry Chanier, Céline Poudat, Benoit Sagot, Georges Antoniadis, Ciara R. Wigham, Linda Hriba, Julien Longhi, and Djamé Seddah. 2014. The comere corpus for french: structuring and annotating heterogeneous cmc genres. Submission to Journal of Language Technology and Computational Linguistics.

M Conover, B Gonçalves, J Ratkiewicz, A Flammini, and F Menczer. 2011. Predicting the political

[6] http://fondation.u-cergy.fr/

alignment of twitter users. In *Proceedings of 3rd IEEE Conference on Social Computing*.

Mats Dahllf. 2012. Automatic prediction of gender, political affiliation, and age in swedish politicians from the wording of their speeches - a comparative study of classifiability. *Literary and Linguistic Computing*, (2):139–153.

Kathleen T. Durant and Michael D. Smith. 2006. Mining sentiment classification from political web logs. In *In Proceedings of Workshop on Web Mining and Web Usage Analysis*.

Kathleen Durant and Michael Smith. 2007. Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection. In *Advances in Web Mining and Web Usage Analysis*, pages 187–206. Springer Berlin / Heidelberg.

Miles Efron. 2006. Using cocitation information to estimate political orientation in web documents. *Knowledge and Information Systems*, (4):492–511.

Emmanuel Giguet and Nadine Lucas. 2001. La détection automatique des citations et des locuteurs dans les textes informatifs.

Swapna Gottipati, Minghui Qiu, Liu Yang, Feida Zhu, and Jing Jiang. 2013. Predicting user's political party using ideological stances. In *Social Informatics*, pages 177–191. Springer.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Association for Computational Linguistics*.

Julien Longhi, Claudia Marinica, Boris Borzic, and Abdul Alkhouli. 2014. Polititweets, corpus de tweets provenant de comptes politiques influents. Technical report.

Julien Longhi. 2008. *Objets discursifs et doxa : essai de sémantique discursive*. L'Harmattan.

Julien Longhi. 2013. Essai de caractérisation du tweet politique. *L'information grammaticale*, pages 125–132.

Julien Longhi. 2014. Le pigeon est-il un canard comme les autres ? esquisse dune théorie des objets discursifs. In *Res Per Nomen IV- Les théories du sens et de la référence - Hommage à Georges Kleiber*. Éditions des Presses Universitaires de Reims.

Ghassan Mourad and Jean-Pierre Desclés. 2003. Identification et extraction automatique des informations citationnelles dans un texte. *Le Discours rapporté dans tous ses états: question de frontieres?*

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In William W. Cohen and Samuel Gosling, editors, *Fourth International AAAI Conference on Weblogs and Social Media*. The AAAI Press.

Fabien Poulard, Thierry Waszak, Nicolas Hernandez, and Patrice Bellot. 2008. Repérage de citations, classification des styles de discours rapporté et identification des constituants citationnels en écrits journalistiques. In *Actes de la 15me Conference sur le Traitement Automatique des Langues Naturelles*.

François Rastier. 2011. *La mesure et le grain. Sémantique de corpus*. Honoré Champion, lettres numriques edition.

Yaroslav Riabinin. 2009. Computational identification of ideology in text: Study of canadian parliamentary debates. Master thesis, University of Toronto.

Georges-Elia Sarfati, 2014. *Les discours institutionnels en confrontation. Contributions a lanalyse des discours institutionnels et politiques*, chapter Lemprise du sens: Note sur les conditions theoriques et les enjeux de lanalyse du discours institutionnel, pages 13–46. LHarmattan.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pages 116–124, Morristown, NJ, USA. Association for Computational Linguistics.

Marilyn A. Walker, Pranav Anand, Rob Abbott, Jean E. Fox Tree, Craig H. Martell, and Joseph King. 2012. That is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, (4):719–729.

Daniel Xiaodan Zhou, Paul Resnick, and Qiaozhu Mei. 2011. Classifying the political leaning of news articles and users from user votes. In *Fifth International AAAI Conference on Weblogs and Social Media*. The AAAI Press.