What does Twitter have to say about ideology?

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

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 politicians 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 will 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 data-sets 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 detection, existing works are based on simple linguistic models as in (Gerrish and Blei, 2011) where the authors predict the voting behavior of legislators on the basis of bag-of-words representations from the proposed bills and deduct legislators' political tendencies. Another type of works use annotated corpus in order to infer lexical characteristics of the ideology; one of these works is (Sim et al., 2013) where authors have used an HMM model (Hiden Marcov Model) to deduct ideologies in candidate discourse during the campaign cycle of united-states in 2012. Similarly, in (Iyyer et al., 2014) the authors introduce 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.

The ideology was defined by multiple authors in multiple occasions. According to Erikson and Tedin in (2003), the ideology is a "...set of beliefs about the proper order of society...". Knight (2006) points out the fact that "Specific ideologies crystallize and communicate the many beliefs, opinions and values of an identifiable group...". This definition is basic, limited to the political camp (right, left, etc.). The ideology refers obviously to the "content" of a discourse, but it can also rely on the "form"; in this context, the discourse analysis field proposes valuable criteria to identify ideology.

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. Then 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; a tweet is ideological if and only if it satisfies all five criteria and all the criteria have the same weight.

- 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.

Below we describe the (linguistic) rules that correspond/implement to each one of the seven criteria. These rules fall 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 declarative criteria. More generally, the theory of discursive objects opens up Sarfati's theory to linguistic corpora.

Criterion 1 is implemented by:

Rule 1: no spatiotemporal 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 is implemented by:

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 is implemented by:

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 is implemented by:

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 is implemented by:

Rule 6 is adequate in order to validate this criterion.

Since a tweet is identified as ideological if and only if it satisfies all the criteria, then, conse-

quently, a tweet has to satisfy all seven rules described above in order to be identified as ideological.

3 Integrating linguistic rules in Natural Language Processing tools

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

Moreover, the implementation of these rules in our system requires a morpho-syntactic analysis in order to determine the part-of-speech category for each word in a tweet: verb, adjective, noun, preposition, etc. For this purpose, we also need to use a suite of NLP tools that carries the corresponding functionality. Thus we reviewed the available open source[2] NLP APIs that we will detail in the following subsection.

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 evaluate their capabilities and limits, and to finally select one or more of them to use. In our study, we are searching for the following elements: verb tenses, adjectives and nouns objective or subjective, personal pronouns, connectors, proper nouns, space and time markers. We tested and evaluated three well-known POS taggers:

- Stanford POS Tagger[3]: offers a Java implementation of the log-linear POS tagger provided by the Stanford NLP group. The provided library allows the user to tag words in the text. The tagger has to load a trained file (named model) containing the necessary information for the tagger. Several trained models are provided by Stanford NLP group for different languages, including French; for French, the model is based on the prelabeled French corpus named Treebank.

- Apache Open NLP[4]: the Apache Open NLP library is a machine learning based toolkit for natural language text processing. 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[5]: 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 according to the Linked Data and the Semantic Web principles. NLGbAse metadata are used to produce resources and training corpora for information extraction tools like Wikimeta. Wikimeta detects named entities, and links them to their RDF description available as Linked Data. 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 also provides TreeTagger, a POS Tagger, and a frequency analysis tool.

In order to compare the POS taggers presented above, we test the performance of their APIs on a set of 100 tweets representing 1920 words. To this end, each API annotates the tweets' words with the corresponding tags, and then we manually compare the results and compute the error rate for each API. The results, presented in Table 1, point out (1) that, regarding the error rate, the Wikimeta Tagger outperforms the other taggers, and (2) that Wikimeta proposes a larger number of tags.

Moreover, the analysis allowed us to determine that, on the one hand, Stanford POS Tagger makes no distinction between nouns and proper

[2]We surveyed only open source APIs both because they are open to anyone to use and the code is available to extend as needed

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

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

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

	Stanford POS Tagger	Apache Open NLP Tagger	Wikimeta Tagger
Error rate	$2,5\%$	$2,55\%$	$2,39\%$
Number of tags	8	13	37

Table 1: Comparison of the results provided by Stanford POS, Apache Open NLP and Wikimeta Taggers.

nouns, between verbs and past participles, and does not tag accordingly verbs' tenses, articles and amounts. On the other hand, Apache Open NLP Tagger does not detect punctuation marks and, as Stanford POS Tagger, does not detect verbs' tenses, articles and amounts although it offers more details than the later.

To conclude, Wikimeta allows us to detect all the elements that we need in order to implement the linguistic rules, such as: verbs' tenses, connectors, proper nouns, personal pronouns. Moreover, it is able to give details concerning proper names, and distinguish between places and people through the detection of named entities (it connects named entities to their RDF description from the linked data).

Based on the results detailed above, we decided 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, using Wikimeta, in our 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 create a set with all these markers and check if they appear in a tweet. For example, now (*maintenant* - fr), tomorrow (*demain* - fr), etc.

Rule 2: Equally, 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. 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 verbs, we first need to find the infinitive form of the verbs in the tweet. To do that, we use a second API[6] that ensures the lemmatization; this API was developed by the Natural Language Processing group of Sheffield University. 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 they exist in the tweet.

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 most dominant verb tense in the tweet. To this end, we count the occurrence 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 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 introduc-

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

tory 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...* - fr);
- the second one is a indirect speech with the form ...explained X (*...a expliqué X* - fr).

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 a significant 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 set of the ideological tweets. A morpho-syntactic analysis is done on the tweets by Wikimeta API allowing POS annotation and detection of named entities. A tweet is identified by the system as ideological only if it satisfies *all* of the seven linguistic rules presented above, knowing that all the rules have the same weight in the system. For each tweet the system notes the rules that it satisfies.

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 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[7]. They use tweets to report their current thoughts and actions, comment on breaking news and even engage in discussions.

4.2 Corpus Description

Nowadays, political tweets are considered by linguistic researchers as a new form of political discourse (Longhi, 2013). Through their tweets, politicians aim to make public their (new) ideas and convictions, but, also to convince the voters that their (the politicians') goals, expectations and actions are the ones to follow and support. In this context, we propose to test our system on a political tweets corpus as there is a bigger probability to contain ideological texts. Moreover doing this, we expect to reduce noise as politicians usually use more standard French when tweeting, avoiding much of web-slang.

The corpus of tweets that we used in our experiments 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

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

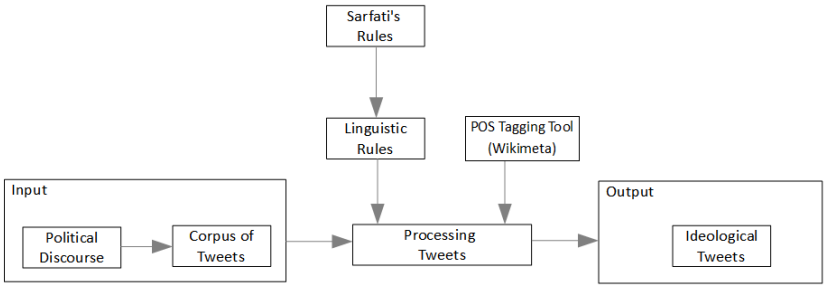

Figure 1: Ideological tweet detection system.

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 published 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. It is important to recall that a tweet is identified as ideological by the system if the tweet satisfies all the 7 rules described above; note that all the 7 rule have the same weight in the system.

Tweet 1: *Je suis ravi de pouvoir compter sur tous ceux qui m'ont accompagné ce soir sur Twitter pendant #motcroises, merci à vous !*

Tweet 2: *Bruno Lemaire : "Les socialistes vivent dans le monde d'avant, c'est pourquoi nous devons inventer le monde d'après."*

Tweet 3: *Le rassemblement ce n'est pas avoir peur les uns des autres, c'est être forts ensemble.*

Tweet 4: *Ns avons perdu ms ns avons gagné un combat: faire naître l'opposition.Le dbut de l'alternance! Merci a chacune et chacun.*

Tweet 1 satisfies Rules 5, 6 and 7, but it does not satisfy Rules 1, 2, 3 and 4: Rule 1 because the tweet contains the word tonight (*ce soir* - fr), Rule 2 as it begins with the interlocution subject I (*je* - fr), Rule 3 because of the presence of the proper noun "Twitter" and Rule 4 as the tweet contains an exclamation mark.

Tweet 2 satisfies Rules 1, 2, 3, 5 and 6, but it does not satisfy Rules 4 and 7: Rule 4 because the tweets contains the modal verb must (*devons* - fr) and Rule 7 as the tweet represents a direct speech where the relator is *Bruno Lemaire* and the speech is between quotes.

Tweet 3 satisfies the 7 rules and is identified as ideological by the system: it does not contain any spatio-temporal marks or proper nouns, interlocution subjects or any connectors, exclamation or interrogation marks, modal verbs or discourse forms; moreover, the verbs' tense is the present.

Tweet 4 satisfies Rules 1, 2, 3, 5, 6 and 7, but it does not satisfy Rule 4. This tweets outlines that web-slangs and abbreviations introduce important issues in our system. Indeed Tweet 4 contains abbreviations for we (*Ns* - *nous* - fr) and for but (*ms* - *mais* - fr) wrongly annotated by Wikimeta. Thus, the system does not detect that Rules 2 and 5 are not satisfied.

However, working on a political tweets corpus ensures us that web-slangs and abbreviations are limited as politicians use proper standard French.

4.4 Results

We tested our system on 20400 tweets selected chronologically from the corpus, and 321 tweets were identified as ideological as they satisfy all 7 rules. Then, we analyzed these results from 3 points of view: (1) the 321 tweets were evaluated in order to compute the precision of our system, (2) the rest of 20079 tweets identified as non-ideological by the system were analyzed in an effort to better understand the recall of our system, and (3) we aimed to detect common linguistic patterns in the ideological tweets.

4.4.1 False positives analysis

The 321 tweets identified as ideological by the system were then manually analyzed for validation by an expert on ideology texts. The purpose

of this analysis is twofold: (1) we wanted to determine how many tweets, from the 321 identified as ideological by the system, are validated as ideological by the expert, and (2) for the tweets that are not validated as ideological by the expert, we expect to identify characteristics that would allow us to refine the results and to distinguish individual traits that can further lead us to improve our system. The result of this analysis is presented in Table 2. From the 321 tweets identified as ideological by the system, 214 tweets are validated as being ideological by the expert representing 66.66% of the 321 tweets. The rest of 33.33% is shared between tweets that are non-ideological and tweets that are partially ideological. In the following, we will detail these two categories.

For the non-ideological tweets, a detailed analyses allowed us to detect the following special cases: (1) a tweet beginning with "@" is usually a response to another tweet and, thus, it is quite brief and not ideological (e.g., *@askolovitchC il faut conduire avec moderation...*); and (2) a tweet containing "#" indicates a very specific context, thus, it cannot be interpreted independently (e.g., *#retraites : visiblement on s'oriente vers du grand n'importe quoi ...*).

The partially ideological tweets are those contextual tweets that can be interpreted out of their context and consequently become ideological. Thus, they have the specificity of allowing two interpretations: ideological and contextual. The following examples describe this type of tweets:

- the tweet *#Confsociale : l'uniformisation et la simplification des systèmes de prévention sociale et de retraite s'impose dès à présent* is contextual as it is related to a specific manifestation. Nevertheless, its content can be clearly understood outside the context.

- the tweet *@DominiqueReynie bravo pour ce travail. l'innovation est forcément une contestation de l'existant* is contextual as its author answers to another tweet, but at the same time he hopes being read by others so he adds an ideological message.

It is important to note that the expert decided to validate as ideological several tweets containing "#" or beginning with "@" as they carry strong ideological messages (e.g., *Le progrès social n'est pas l'adversaire de la performance économique #loiESS*).

4.4.2 False negatives analysis

After analyzing the set of tweets identified as ideological by the system, we also analyzed the set of tweets identified as non-ideological by the system with the aim to determine if ideological tweets have been misclassified by our system as non-ideological.

To this end, we sampled the set of tweets identified as non-ideological by the system (20079 tweets) by randomly selecting 4% of the tweets that do not satisfy only one rule (117 tweets) and 2% of the tweets falling in the other categories (329 tweets). Thus, we obtained a set of 446 tweets that was analyzed for validation by the expert. This analysis showed that 96.64% of the sampled tweets were classified correctly as non ideological, thus leaving the false negatives to represent 3.36%. One other observation is that there were no errors if a tweet does not satisfy 3 rules or more; this tweet is always correctly identified by the system as non-ideological.

Furthermore, in order to understand why these tweets were misclassified by the system, we carefully analyzed the false negatives and we made the following conclusions: (1) several misclassifications result as an error of annotation of Wikimeta; (2) several misclassifications are caused by Rule 2 as sometimes interlocution subjects (as our, *nos* - fr) are used as general referent; and (3) Rule 6 produces some misclassifications equally when the future tense dominating the tweet is prospective (e.g., *La République sera à tous les Français*). These observations will be exploited to further improve the system's performance in the future.

4.4.3 Linguistic structures identification

Analyzing the ideological tweets, the expert pointed out that they contain a style that fits into a rhetorical and strongly argumentative reference in order to give them more strength and to impose the ideology.

In this context, some structures were clearly identified:

Have to (*Il faut* - fr): e.g., *Ce qu'il faut c'est établir des priorités, choisir des filières*

Expert validation of the 321 tweets identified as ideological by the system		
Ideological tweets	Non-ideological tweets	Partially ideological tweets
214 (66.66%)	75 (23.36%)	32 (9.96%)

Table 2: Results after expert's validation of the 321 tweets identified as ideological by the system.

d'excellence, créer des emplois dans des secteurs porteurs.

There is (*Il y a* - fr): e.g., *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...is x... or which is...that is... (*X, c'est x* or *ce qui est...c'est* - fr): e.g., *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*

At the same time, the expert observed that the current hypothesis of detecting ideological tweets can be enriched with style-based criteria, which could give interesting results.

Furthermore, regarding Rule 4, it might be interesting to evaluate the tweets containing the have to verb (*devoir* - fr), as in some cases the verb have to does not necessarily indicates the involvement of the speaker, but rather a form of general truth, e.g., *Les démocrates doivent s'unir pour mettre fin à cette violence dans le débat public. #BFMTV*.

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*.

5 Conclusions and Future Work

In this paper, we implemented Sarfati's criteria as a set of linguistic rules for detecting ideology in textual documents. Moreover, we developed a system that implements these rules as an extension of an NLP System. Finally, we tested our system against a set of 20400 tweets of French politicians in order to experiment rules' implementation and their accuracy.

The evaluation of the rules and their implementation give us good results for the system's accuracy since 66.66% of tweets identified as ideological were indeed so and 96.64% of tweets identified as non-ideological (after sampling) were validated as non-ideological by the expert.

For the future work, we plan to take advantage of the analysis produced by the expert in order to revise or relax some of the rules that might misclassify some tweets, but also to propose a set of rules allowing us to detect the type of the ideology for those ideological tweets. Moreover, we plan 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[8]) 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 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

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

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.

RS Erikson and KL Tedin. 2003. *American Public Opinion*. Longman.

Sean Gerrish and David M. Blei. 2011. Predicting legislative roll calls from text. In *International Conference on Machine Learning (ICML)*, pages 489–496.

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*.

Kathaleen Knight. 2006. Transformations of the concept of ideology in the twentieth century. *American Political Science Review*, pages 619–626.

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.

Yanchuan Sim, Brice D. L. Acree, Justin H. Gross, and Noah A. Smith. 2013. Measuring ideological proportions in political speeches. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing and Natural Language Learning*.

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.