Syracuse University
School of Information Studies
IST 565

School of Information Studies
Syracuse University

IST565 DATA MINING
Final Project Report

Instructor: Bei Yu

Student: David Forteguerre
SUID: 673608207
Email: dfortegu@syr.edu

Academic Year 2017 – 2018

I. INTRODUCTION

For this final project, I will explore the different levels of formality and registers of the French language using emails as the main form of communication.
English and Romance languages work very differently when it comes to addressing someone. In English, the pronoun *you* is the only option that can be used, whereas in Romance languages, there is a clear distinction in pronoun usage as someone may be addressed either informally (i.e., with the pronoun *tu* in most Romance languages) or formally (e.g. with the pronoun *vous* in French, *usted* in Spanish, *Lei* in Italian, etc.). Thus, the choice of pronoun in Romance languages is already a good indicator of the degree of formality of a given situation, even though it is not the only one. In France, a very high degree of formality is expected in many day-to-day situations where other countries and cultures would typically be more informal. For instance, in Spain, it is common to address a professor using *tu* instead of *usted*, to be on a first-name basis, and to greet each other saying *Hola*! (Hi!). In France, this type of behavior is unacceptable and would be seen as highly disrespectful. French professors may never be addressed with *tu* or by their first name; *vous* is the norm. Greetings also need to be more formal, so *Bonjour*! (Good morning!) would be used instead of *Salut*! (Hi!). The discrepancy in the level of formality required in different European countries for the same kind of situation can be explained by the fact that French culture has always shown a strong tradition of high respect and politeness, which has always been reflected in the language. This is certainly linked to the history of the language and also to the heavy influence of the French Academy, an institution that heavily regulates and protects the French language.
As a result, French is certainly the best language to choose for this analysis, culturally and linguistically speaking.

The purpose of this project is to discover the main linguistic patterns in the levels of formality of French emails. This is a typical classification task and several algorithms will be used in order to solve it. The classifiers will learn those patterns and ultimately output the main indicators of what makes an email either formal or informal.

Even though several data validation techniques will be used to evaluate the accuracy of the results (e.g. cross validation), I will also use my knowledge of French as a native speaker to evaluate the findings. The goal of this project is more educational than exploratory. Indeed, I plan to demonstrate a linguistic and sociolinguistic phenomenon through the use of algorithms to ultimately be able to explain it with real scientific evidence to a non-native speaker. From the beginning of the project, I already had a clear idea of what the algorithms were going to output, and my findings were eventually confirmed.

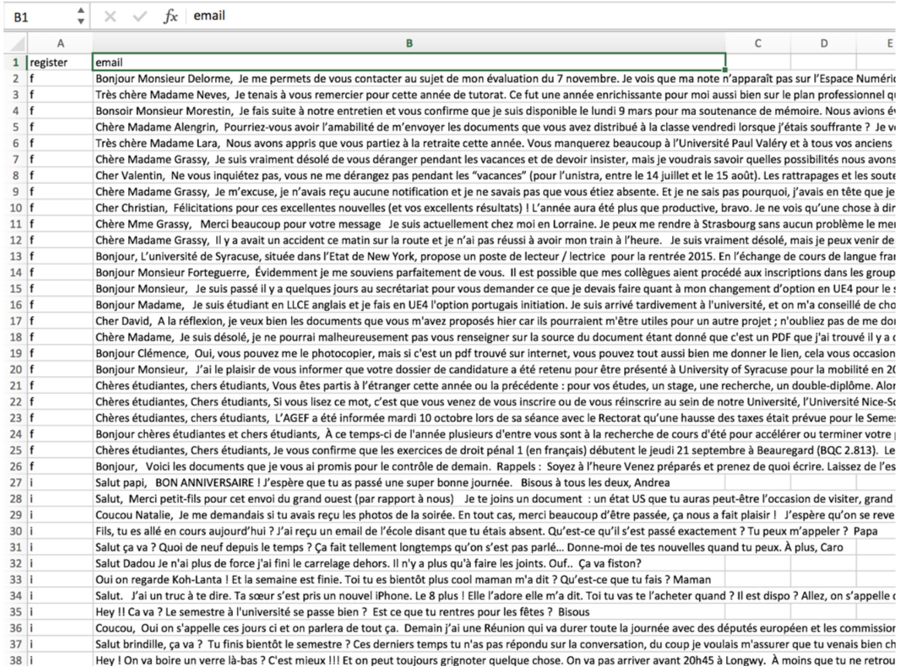
II. DATA COLLECTION

The data collection was the most tedious and time-consuming part of this project, but I wanted to make sure that I would have high-quality emails. When doing language analysis, there are key considerations to always have. The data must always reflect how the language is actually used by a variety of people. Thus, I decided to ask my friends, their friends, as well as my family members to send me email samples. I made sure to ask people from all across France (mostly Southern France, Northern France, and Eastern France), and to ask people from different age ranges. I also included some of the emails I sent to my French 201 students at Syracuse University last year.

I collected **50** long and complete formal emails (more specifically university-related emails) and **50** informal emails (daily communications between family and friends.) As you know, the more data, the more accurate the algorithms output. Although it is often necessary to have a large amount of data for this type of task, I found the results to already be significant and very conclusive after running my analysis on 50 emails and then 100.

Even though a language often shows several registers and levels of formality (e.g. very formal, formal, everyday French, informal, slang), I decided to focus on the two major ones and create a data frame categorizing my data into binary bins: f (*formal*) vs. i (*informal*).

Below is a screenshot with fewer emails that demonstrates the organization of my data in Microsoft Excel.



B1 *fx* email

	A	B	C	D	E
1	register	email			
2	f	Bonjour Monsieur Delorme, Je me permets de vous contacter au sujet de mon évaluation du 7 novembre. Je vois que ma note n’apparaît pas sur l’Espace Numéri			
3	f	Très chère Madame Neves, Je tenais à vous remercier pour cette année de tutorat. Ce fut une année enrichissante pour moi aussi bien sur le plan professionnel q			
4	f	Bonsoir Monsieur Morestin, Je fais suite à notre entretien et vous confirme que je suis disponible le lundi 9 mars pour ma soutenance de mémoire. Nous avions é			
5	f	Chère Madame Alengrin, Pourriez-vous avoir l’amabilité de m’envoyer les documents que vous avez distribué à la classe vendredi lorsque j’étais souffrante ? Je v			
6	f	Très chère Madame Lara, Nous avons appris que vous partiez à la retraite cette année. Vous manquerez beaucoup à l’Université Paul Valéry et à tous vos anciens			
7	f	Chère Madame Grassy, Je suis vraiment désolé de vous déranger pendant les vacances et de devoir insister, mais je voudrais savoir quelles possibilités nous avons			
8	f	Cher Valentin, Ne vous inquiétez pas, vous ne me dérangez pas pendant les “vacances” (pour l’unistra, entre le 14 juillet et le 15 août). Les rattrapages et les soute			
9	f	Chère Madame Grassy, Je m’excuse, je n’avais reçu aucune notification et je ne savais pas que vous étiez absente. Et je ne sais pas pourquoi, j’avais en tête que je			
10	f	Cher Christian, Félicitations pour ces excellentes nouvelles (et vos excellents résultats) ! L’année aura été plus que productive, bravo. Je ne vois qu’une chose à dir			
11	f	Chère Mme Grassy, Merci beaucoup pour votre message Je suis actuellement chez moi en Lorraine. Je peux me rendre à Strasbourg sans aucun problème le me			
12	f	Chère Madame Grassy, Il y a avait un accident ce matin sur la route et je n’ai pas réussi à avoir mon train à l’heure. Je suis vraiment désolé, mais je peux venir de			
13	f	Bonjour, L’université de Syracuse, située dans l’Etat de New York, propose un poste de lecteur / lectrice pour la rentrée 2015. En l’échange de cours de langue fra			
14	f	Bonjour Monsieur Forteguerre, Évidemment je me souviens parfaitement de vous. Il est possible que mes collègues aient procédé aux inscriptions dans les group			
15	f	Bonjour Monsieur, Je suis passé il y a quelques jours au secrétariat pour vous demander ce que je devais faire quant à mon changement d’option en UE4 pour le			
16	f	Bonjour Madame, Je suis étudiant en LLCE anglais et je fais en UE4 l'option portugais initiation. Je suis arrivé tardivement à l'université, et on m'a conseillé de cho			
17	f	Cher David, A la réflexion, je veux bien les documents que vous m'avez proposés hier car ils pourraient m'être utiles pour un autre projet ; n'oubliez pas de me do			
18	f	Chère Madame, Je suis désolé, je ne pourrai malheureusement pas vous renseigner sur la source du document étant donné que c'est un PDF que j'ai trouvé il y a			
19	f	Bonjour Clémence, Oui, vous pouvez me le photocopier, mais si c'est un pdf trouvé sur internet, vous pouvez tout aussi bien me donner le lien, cela vous occasion			
20	f	Bonjour Monsieur, J’ai le plaisir de vous informer que votre dossier de candidature a été retenu pour être présenté à University of Syracuse pour la mobilité en 2			
21	f	Chères étudiantes, chers étudiants, Vous êtes partis à l’étranger cette année ou la précédente : pour vos études, un stage, une recherche, un double-diplôme. Alor			
22	f	Chères étudiantes, Chers étudiants, Si vous lisez ce mot, c’est que vous venez de vous inscrire ou de vous réinscrire au sein de notre Université, l’Université Nice-S			
23	f	Chères étudiantes, chers étudiants, L’AGEF a été informée mardi 10 octobre lors de sa séance avec le Rectorat qu’une hausse des taxes était prévue pour le Seme			
24	f	Bonjour chères étudiantes et chers étudiants, À ce temps-ci de l'année plusieurs d'entre vous sont à la recherche de cours d'été pour accélérer ou terminer votre			
25	f	Chères étudiantes, Chers étudiants, Je vous confirme que les exercices de droit pénal 1 (en français) débutent le jeudi 21 septembre à Beauregard (BQC 2.813). Le			
26	f	Bonjour, Voici les documents que je vous ai promis pour le contrôle de demain. Rappels : Soyez à l’heure Venez préparés et prenez de quoi écrire. Laissez de l’es			
27	i	Salut papi, BON ANNIVERSAIRE ! J’espère que tu as passé une super bonne journée. Bisous à tous les deux, Andrea			
28	i	Salut, Merci petit-fils pour cet envoi du grand ouest (par rapport à nous) Je te joins un document : un état US que tu auras peut-être l’occasion de visiter, grand			
29	i	Coucou Natalie, Je me demandais si tu avais reçu les photos de la soirée. En tout cas, merci beaucoup d’être passée, ça nous a fait plaisir ! J’espère qu’on se reve			
30	i	Fils, tu es allé en cours aujourd’hui ? J’ai reçu un email de l’école disant que tu étais absent. Qu’est-ce qu’il s’est passé exactement ? Tu peux m’appeler ? Papa			
31	i	Salut ça va ? Quoi de neuf depuis le temps ? Ça fait tellement longtemps qu’on s’est pas parlé… Donne-moi de tes nouvelles quand tu peux. À plus, Caro			
32	i	Salut Dadou Je n'ai plus de force j'ai fini le carrelage dehors. Il n'y a plus qu'à faire les joints. Ouf.. Ça va fiston?			
33	i	Oui on regarde Koh-Lanta ! Et la semaine est finie. Toi tu es bientôt plus cool maman m'a dit ? Qu’est-ce que tu fais ? Maman			
34	i	Salut. J’ai un truc à te dire. Ta sœur s’est pris un nouvel iPhone. Le 8 plus ! Elle l’adore elle m’a dit. Toi tu vas te l’acheter quand ? Il est dispo ? Allez, on s’appelle			
35	i	Hey !! Ca va ? Le semestre à l'université se passe bien ? Est ce que tu rentres pour les fêtes ? Bisous			
36	i	Coucou, Oui on s'appelle ces jours ci et on parlera de tout ça. Demain j’ai une Réunion qui va durer toute la journée avec des députés européen et les commissior			
37	i	Salut brindille, ça va ? Tu finis bientôt le semestre ? Ces derniers temps tu n'as pas répondu sur la conversation, du coup je voulais m'assurer que tu venais bien ch			
38	i	Hey ! On va boire un verre là-bas ? C'est mieux !!! Et on peut toujours grignoter quelque chose. On va pas arriver avant 20h45 à Longwy. À moins que tu ne retrou			

emailsfrdata.csv[1]

Considering how difficult it can be to process data in Weka and following Professor Yu’s recommendation, I decided to manually create the .arff file I would later import into Weka instead of attempting to convert my .csv file to an .arff file automatically.

However, in order to create that .arff file, I had to perform some preliminary steps on the data:

- I had to remove last names at the end of each email to preserve confidentiality.

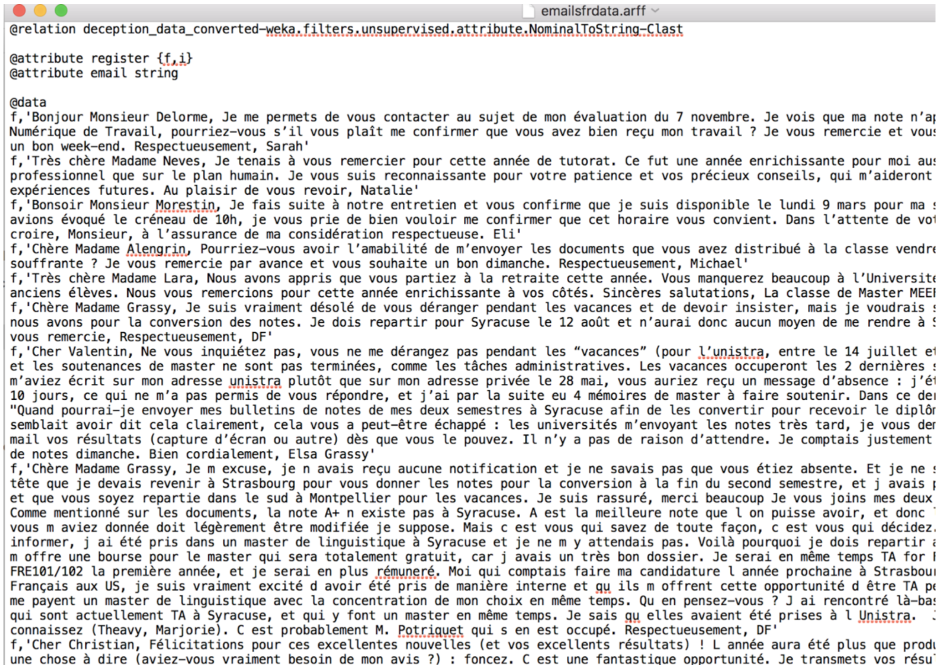
[1] From just looking at the data, we can already notice some patterns at the beginning of the emails.

- I had to remove line breaks from the data to make sure all emails where in a paragraph form without line breaks.
- I also had to remove all apostrophes from the data. Apostrophes are very common in French as many grammatical words tend to be abbreviated when they come before a word that starts with a vowel. Oftentimes, those abbreviations are a grammatical requirement and not a register choice, unlike in English.

Abbreviations: **English vs. French**

	1st Element +	2nd Element	Abbreviation	Required?
French	je (*I*)	ai (*have*)	j'ai (*I have*)	Required, as *je ai* would be grammatically incorrect
English	I	am	I'm	Not required. *I am* can also be used as a more formal option

Only then was I able to create the .arff file. Below is a preview of the document.



```
@relation deception_data_converted-weka.filters.unsupervised.attribute.NominalToString-Clast

@attribute register {f,i}
@attribute email string

@data
f,'Bonjour Monsieur Delorme, Je me permets de vous contacter au sujet de mon évaluation du 7 novembre. Je vois que ma note n'ap
Numérique de Travail, pourriez-vous s'il vous plaît me confirmer que vous avez bien reçu mon travail ? Je vous remercie et vous
un bon week-end. Respectueusement, Sarah'
f,'Très chère Madame Neves, Je tenais à vous remercier pour cette année de tutorat. Ce fut une année enrichissante pour moi aus
professionnel que sur le plan humain. Je vous suis reconnaissante pour votre patience et vos précieux conseils, qui m'aideront
expériences futures. Au plaisir de vous revoir, Natalie'
f,'Bonsoir Monsieur Morestin, Je fais suite à notre entretien et vous confirme que je suis disponible le lundi 9 mars pour ma s
avions évoqué le créneau de 10h, je vous prie de bien vouloir me confirmer que cet horaire vous convient. Dans l'attente de vot
croire, Monsieur, à l'assurance de ma considération respectueuse. Eli'
f,'Chère Madame Alengrin, Pourriez-vous avoir l'amabilité de m'envoyer les documents que vous avez distribué à la classe vendre
souffrante ? Je vous remercie par avance et vous souhaite un bon dimanche. Respectueusement, Michael'
f,'Très chère Madame Lara, Nous avons appris que vous partiez à la retraite cette année. Vous manquerez beaucoup à l'Universite
anciens élèves. Nous vous remercions pour cette année enrichissante à vos côtés. Sincères salutations, La classe de Master MEEF
f,'Chère Madame Grassy, Je suis vraiment désolé de vous déranger pendant les vacances et de devoir insister, mais je voudrais s
nous avons pour la conversion des notes. Je dois repartir pour Syracuse le 12 août et n'aurai donc aucun moyen de me rendre à S
vous remercie, Respectueusement, DF'
f,'Cher Valentin, Ne vous inquiétez pas, vous ne me dérangez pas pendant les "vacances" (pour l'unistra, entre le 14 juillet et
et les soutenances de master ne sont pas terminées, comme les tâches administratives. Les vacances occuperont les 2 dernières s
m'aviez écrit sur mon adresse unistra plutôt que sur mon adresse privée le 28 mai, vous auriez reçu un message d'absence : j'ét
10 jours, ce qui ne m'a pas permis de vous répondre, et j'ai par la suite eu 4 mémoires de master à faire soutenir. Dans ce der
"Quand pourrai-je envoyer mes bulletins de notes de mes deux semestres à Syracuse afin de les convertir pour recevoir le diplôm
semblait avoir dit cela clairement, cela vous a peut-être échappé : les universités m'envoyant les notes très tard, je vous der
mail vos résultats (capture d'écran ou autre) dès que vous le pouvez. Il n'y a pas de raison d'attendre. Je comptais justement
de notes dimanche. Bien cordialement, Elsa Grassy'
f,'Chère Madame Grassy, Je m excuse, je n avais reçu aucune notification et je ne savais pas que vous étiez absente. Et je ne s
tête que je devais revenir à Strasbourg pour vous donner les notes pour la conversion à la fin du second semestre, et j avais p
et que vous soyez repartie dans le sud à Montpellier pour les vacances. Je suis rassuré, merci beaucoup Je vous joins mes deux
Comme mentionné sur les documents, la note A+ n existe pas à Syracuse. A est la meilleure note que l on puisse avoir, et donc l
vous m aviez donnée doit légèrement être modifiée je suppose. Mais c est vous qui savez de toute façon, c est vous qui décidez.
informer, j ai été pris dans un master de linguistique à Syracuse et je ne m y attendais pas. Voilà pourquoi je dois repartir a
m offre une bourse pour le master qui sera totalement gratuit, car j avais un très bon dossier. Je serai en même temps TA for F
FRE101/102 la première année, et je serai en plus rémuneré. Moi qui comptais faire ma candidature l année prochaine à Strasbour
Français aux US, je suis vraiment excité d avoir été pris de manière interne et qu ils m offrent cette opportunité d être TA pe
me payent un master de linguistique avec la concentration de mon choix en même temps. Qu en pensez-vous ? J ai rencontré là-bas
qui sont actuellement TA à Syracuse, et qui y font un master en même temps. Je sais qu elles avaient été prises à l Unistra.  J
connaissez (Theavy, Marjorie). C est probablement M. Potriquet qui s en est occupé. Respectueusement, DF'
f,'Cher Christian, Félicitations pour ces excellentes nouvelles (et vos excellents résultats) ! L année aura été plus que produ
une chose à dire (aviez-vous vraiment besoin de mon avis ?) : foncez. C est une fantastique opportunité. Je transmets vos résul
```

emailsfrdata.arff

III. DATA PREPROCESSING *(in Weka)*

Even though the dataset was now clean and in the right format, the data still needed to be preprocessed. Thus, I imported the emailsfrdata.arff file into Weka, and then selected the StringToWordVector filter (Unsupervised>attribute>StringToWordVector) in the "preprocess" tab. In the filter settings, I tuned some key parameters:

- I set **attributeIndices** to 2 (→ to apply the filter to the email variable").
- I set **lowerCaseTokens** to "true" (→ to make all capital letters lower case).

- I made sure that **minTermFrequencies** was set to “1” (default) (→ to remove any word that had only been used only *once* in all the emails combined).
- I set **normalizeDocLength** to normalize all data (→ to be able to see the word counts by clicking on each word after applying the filter).
- I set **stopwordsHandler** to my own stopwords file which I created (see below*). The file is called ***stopwordsfr1.txt***.
- I set **tokenizer** to WordDelimiter (--> a basic tokenizer for this data).
- I set **wordsToKeep** to 2000 (instead of 1000, the default) (→ to have more data).

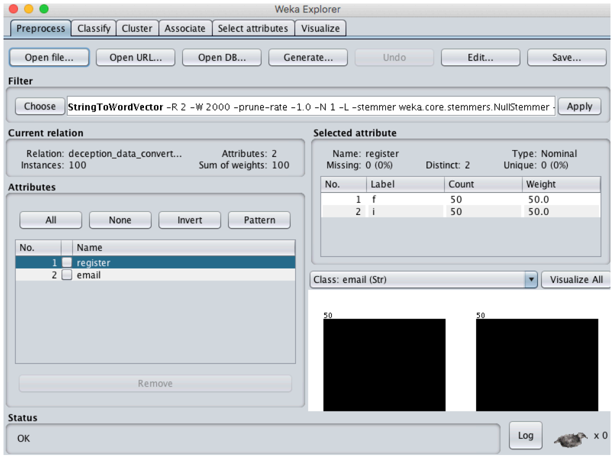
Finally, I applied the filter to the email variable whose content was processed and converted into word tokens.

*Stop words are words that are filtered out before or after processing of natural language data (text). Usually, these words are the most common non-content-bearing words in a language. However, there are no set stop words lists, as the list could vary depending on the nature of the analysis. In my case, I used R to get a list of French stop words. Below is my code:

```
# To get a list of French stopwords
library(tm)
FrenchStopWords <- stopwords("french")
print(FrenchStopWords, quote=FALSE)
```

I exported this list as a .txt file, made sure that each word was on a separate line, and finally went through all the words to make sure I would not be deleting any crucial words from the analysis. After going through the list, I found that “vous” (and its derivatives) and “tu” (and its derivatives) were in the list. Thus, I put a # sign before those words to make sure they would be kept in the dataset.

Below is a screenshot *before* applying the filter, and the output generated *after* applying the filter.



```
StringToWordVector -R 2 -W 2000 -prune-rate -1.0 -N 1 -L -stemmer weka.core.stemmers.NullStemmer -
```

Relation: deception_data_convert...   Attributes: 2
Instances: 100   Sum of weights: 100

Name: register   Type: Nominal
Missing: 0 (0%)   Distinct: 2   Unique: 0 (0%)

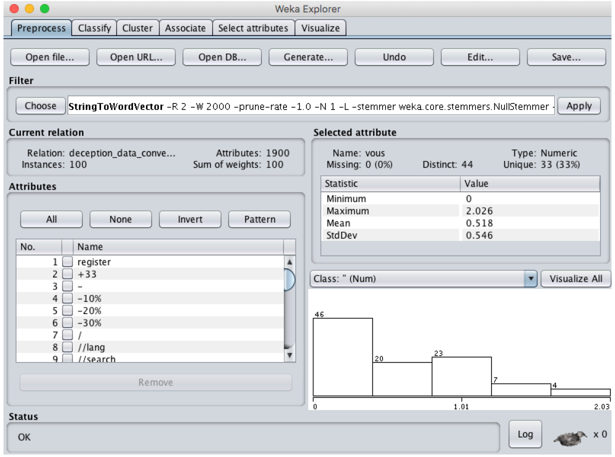
No.	Label	Count	Weight
1	f	50	50.0
2	i	50	50.0

No.		Name
1	☐	register
2	☐	email

Class: email (Str)

OK

Before



```
StringToWordVector -R 2 -W 2000 -prune-rate -1.0 -N 1 -L -stemmer weka.core.stemmers.NullStemmer -
```

Relation: deception_data_conve...   Attributes: 1900
Instances: 100   Sum of weights: 100

Name: vous   Type: Numeric
Missing: 0 (0%)   Distinct: 44   Unique: 33 (33%)

Statistic	Value
Minimum	0
Maximum	2.026
Mean	0.518
StdDev	0.546

No.		Name
1	☐	register
2	☐	+33
3	☐	-
4	☐	-10%
5	☐	-20%
6	☐	-30%
7	☐	/
8	☐	//lang
9	☐	//search

Class: '' (Num)

OK

After

IV. DATA ANALYSIS *(in Weka)*

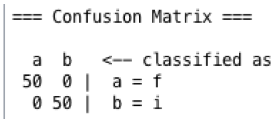
In data mining, a very common technique is to run several algorithms on the same dataset and tune their parameters in order to see which one works best and outputs the highest accuracy. Indeed, each algorithm has unique properties and tends to be best suited for a specific set of tasks only.
For this classification task, I decided to run three algorithms: (1) Naïve Bayes, (2) Support Vector Machines, and (3) Decision Trees in order to see which would work best.

1. Naïve Bayes[2]

I first ran the Naïve Bayes Multinomial algorithm on the dataset.[3] The data was ready to be analyzed, thus I selected the NaiveBayesMultinomial algorithm in the "classify tab", and made sure that the target variable was set to "register" (i.e. what we are trying to predict.) I used 10-fold cross validation in order to evaluate the model.[4]
The results are as follows:

Naïve Bayes algorithm	
Default settings:	**Correctly Classified Instances 100 100 %** Observation: all instances have been classified accurately.

Let's now look at the confusion matrix for the result. It shows that all instances have been classified accurately.



```
=== Confusion Matrix ===

  a  b   <-- classified as
 50  0 |  a = f
  0 50 |  b = i
```

This perfect accuracy may point out that there are some obvious patterns in the data that make it very easy for the algorithm to find the answer. By taking a look at the probabilities of given words, and keeping in mind how pronouns work in French, we can observe major differences. Let's take the example of "tu" (informal *you*).

"tu"		
	f	**i**
Probability output	3.982777131170316E-4	0.01294848936653945
Value rounded-up	**0.0003**	**0.012**

[2] Note that several Naïve Bayes algorithms exist. NB Multinomial is the one that is particularly designed for text categorization.

[3] There are two NB Multinomial algorithms in Weka. One is called NaiveBayesMultinomial and needs the data to be preprocessed and filtered as input, and the other is called NaiveBayesMultinomialText—it actually takes care of the preprocessing and filtering section. I tried both algorithms. For NaiveBayesMultinomialText, I had to click "Undo" in the preprocess tab to undo the output of my StringToWordVector, and then tune the parameters within the algorithm in a similar way before running it. **NaiveBayesMultinomial** actually showed a much higher accuracy. Thus, NaiveBayesMultinomialText was disregarded.

[4] Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples.

As you can see, the informal pronoun “tu” is much more likely to appear in an informal email than in a formal one. So far, our hypothesis has been verified. Let’s go even further and use a different algorithm.

2. SVMs

The second step was to run Support Vector machines on the dataset. Thus, I selected the SMO algorithm (under “functions”) and made sure that the target variable was still set to “register.” Just like for NB, I used 10-fold cross validation in order to evaluate the model.
The results are as follows:

SVM algorithm	
Default settings:	Default kernel (= PolyKernel) **Correctly Classified Instances 100 100 %** Observation: all instances have been classified accurately.

Once again, all emails have been classified accurately. The confusion matrix was the same as the NB algorithm matrix. Using another algorithm and getting similar results really emphasizes that there must be something in the data that betrays the level of formality of each email. It was not even necessary to tune the algorithm parameters for this specific task to try and get a higher accuracy, as the outputs were already 100% accurate (which is unusual.)
Once again, let’s take a look at the attribute weights to see if we can discover any patterns correlated with the usage of pronouns in French.

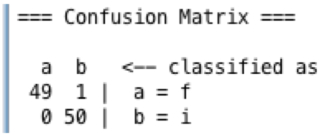
Pronoun	**Weight**
“vous”	+ -0.3534 * (normalized) vous
“tu”	+ 0.3058 * (normalized) tu
Note that a positive weight predicts an informal email in this output, whereas a negative weight predicts a formal email.	

3. Decision Trees

Finally, I decided to run the Decision Tree algorithm. My thought process was that this type of algorithm would be very well suited for this type of task where only a few key elements seem to play a determining role in our predictions. Indeed, it is possible to easily visualize the patterns found by this algorithm thanks to an actual tree output.
To build this model, I selected Weka’s J48 algorithm and made sure the target variable was still set to “register”. Once again, I used 10-fold cross validation in order to evaluate the model.
The results are as follows:

J48 algorithm	
Default settings:	**Correctly Classified Instances 99 99 %** Observation: once again, the accuracy obtained is exceptionally high.

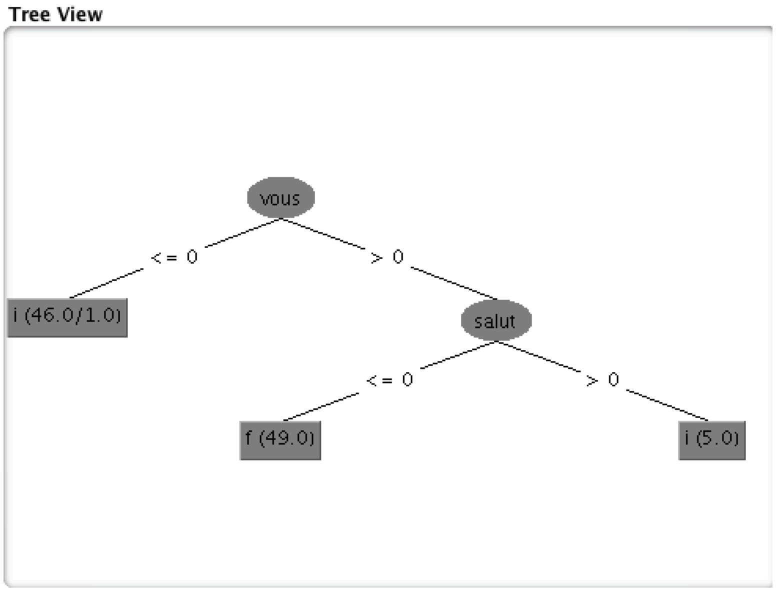
Below is the confusion matrix of the output. We can see that a formal email has been mistakenly classified as informal.



```
=== Confusion Matrix ===

  a  b   <-- classified as
 49  1 |  a = f
  0 50 |  b = i
```

Here is the visualization of the tree generated by the model:



This tree confirms the hypothesis introduced at the beginning of this project. “Vous” (formal *you*) seems to be *by far* the linguistic element that betrays the level of formality of French emails.

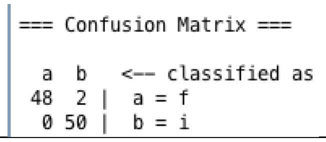
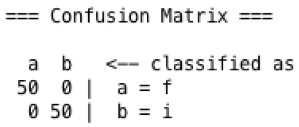
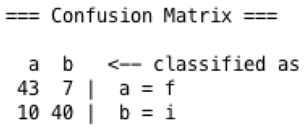
This tree suggests that if there is no “vous” in a given email, it is likely to be informal.[5] On the other hand, if there is one or more “vous”, then the email could very well be formal or not. That is when the analysis becomes even more interesting and informative. The second key indicator in the data I collected seems to be the informal greeting “salut” (*hi*). If an email contains one or more “vous” *and* contains one or more “salut”, then it is likely to be informal. Indeed, this greeting is typically used in very informal situations and would never be used in a formal email (such as a university-related communication). However, if an email contains one or more “vous” and *does not* contain any “salut”, then it is very likely to be formal.

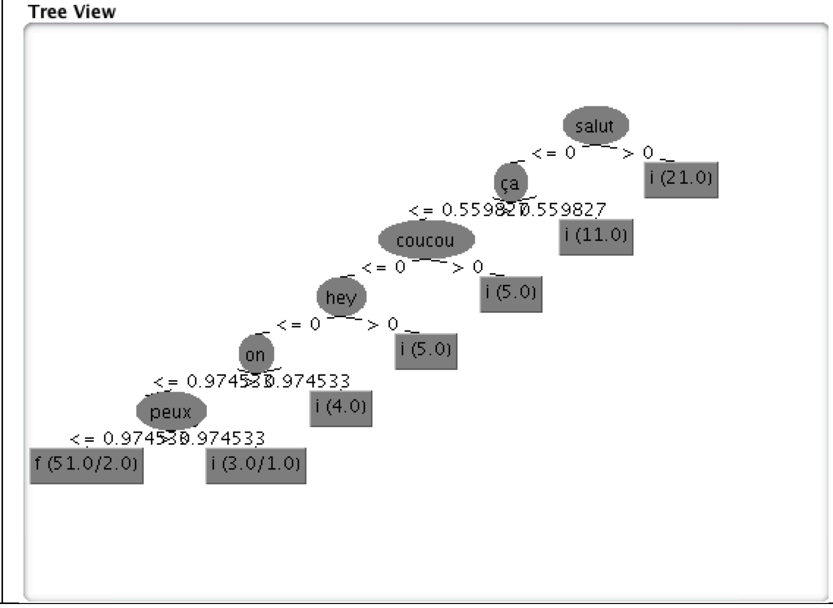
Conclusion

[5] Note that the error found must come from the fact that “vous” is not only used to mean *you* (formal, to address 1 person), but also *you* (no specific register, to address several people) (= “you guys”).

The algorithms used were all helpful in this analysis. As suggested, pronoun usage seems to be key in French to determine the register of a given communication.

However, I decided to push this analysis further and take out “tu” and “vous” from the data to see how the algorithms would perform and discover what the next patterns would be. In order to do so, I created a second stop words list (***stopwordsfr2.txt***) and removed the # signs I had placed before “vous” (and its derivatives) and “tu” (and its derivatives). I then reimported the dataset into Weka, and reapplied the StringToWordVector filter, this time selecting the brand new stop words list. I double-checked to make sure that the pronouns “vous” and “tu” had been removed from the data, and then ran the three algorithms again, using 10-fold cross validation.
Below is a quick summary of the results.

<table>
<tr><td><b>1. NB</b></td><td><b>Correctly Classified Instances 98 98 %</b>
<b>Incorrectly Classified Instances 2 2 %</b>

=== Confusion Matrix ===
a b <-- classified as
48 2 | a = f
0 50 | b = i</td></tr>
<tr><td><b>2. SVMs</b></td><td><b>Correctly Classified Instances 100 100 %</b>

=== Confusion Matrix ===
a b <-- classified as
50 0 | a = f
0 50 | b = i</td></tr>
<tr><td><b>3. Decision Trees</b></td><td><b>Correctly Classified Instances 83 83 %</b>
<b>Incorrectly Classified Instances 17 17 %</b>

=== Confusion Matrix ===
a b <-- classified as
43 7 | a = f
10 40 | b = i</td></tr>
</table>



Running the algorithms again without *tu* and *vous* tells us a lot about the data and the performance of the algorithms to solve this type of task. Indeed, it is now becoming obvious that SVMs seem to be best suited for the task, NB coming in second position, and decision trees coming last. However, we also notice that the accuracy given by decision trees is 83%. Even though it is now much lower, it remains a very high accuracy overall.

This extra step showed us that some other key words also play a role in determining register in French, and these are:

1. ***Salut*** (also found in the previous results), which means *hi* and is typically used in informal contexts.
2. ***Ça***, which means *this* or *that*. However, *ça* is the informal abbreviation of *cela*. Thus, this demonstrative is used in informal contexts as well and serves as the second best indicator.
3. ***Coucou***, which means *hi*. It is another informal greeting.
4. ***Hey***, which also means *hi/hey.* Again, it is an informal greeting.
5. ***On***, which is the informal “nous” (i.e. the informal *we*). This shows that another pronoun (other than “vous” and “tu”) is also a very good indicator.
6. […] Then, we start getting a list of inflected verbs and conjugations (such as “peux”), which is not relevant to this analysis.

As a final step, I decided to use GainRatio on the data (including “tu” and “vous” again) and list the top-20 features. In order to do so, I selected the InfoGainAttributeEval evaluator in the “select attribute” tab. Below is the list of the rankings:

0.8601	1829 **vous**
0.4142	1739 **tu**
0.2508	1538 **salut**

0.2461	1081	ne
0.2266	1824	votre
0.1953	980	madame
0.1853	272	bonjour
0.1821	1050	monsieur
0.1821	1853	**ça**
0.1692	1139	on
0.1441	450	cordialement
0.1441	77	a
0.1441	1433	remercie
0.1198	1823	vos
0.108	1264	plus
0.108	266	bisous
0.108	1673	te
0.108	1771	va
0.0964	704	faire
0.0964	451	**coucou**

This list confirms our findings. Pronouns and greetings played a key role in this analysis, and those features helped our classifiers learn different patterns in the French email data.

V. CONCLUSIONS

To conclude, we have seen that our algorithms have successfully carried out this text classification task.

- First of all, we have seen that the difference of pronoun usage between "tu" and "vous" plays a key role in determining the level of formality of French emails. However, those pronouns were not the only factors, and greetings came next. Indeed, some greetings such as "salut" (*hi*!) are only found in very informal contexts, and also play a role in betraying the register of a given email.
- This analysis also emphasizes a very important difference between French (and more generally, Romance Languages) and English. Romance languages tend to have a wider variety of pronouns, and each pronoun usually conveys very specific ideas and information. Indeed, there is no distinction between *you* (singular) and *you* (plural) in English, which is why this type of analysis would be irrelevant to the English language.
- Finally, this analysis not only highlights two linguistic patterns, but also emphasizes a major difference between American and French cultures. If French culture was as informal and relaxed as American culture (where students often address their professors with an informal tone and often write to them using informal greetings such as "hi" or "hello"), this analysis would not have been as conclusive. Thus, we also learned that the education system in France always requires a high level of formality.

Real life application

As we learned that French culture is much stricter than American culture when it comes to formality, one could imagine a situation where a French university wants to create and implement its own focused inbox that would filter out any email that is not work-related. Even though this concept already exists (Microsoft Outlook does offer the option to have a focused inbox to only keep important emails), it is unlikely that the algorithms developed by Microsoft only look at formality levels in order to classify important and trivial emails due to the informal nature of American culture. Indeed, an important email (e.g., a professor's reminder) may still have an informal tone in the U.S. This would not be the case in France, and it would be much easier to implement this type of tool using formality as the main (and only?) factor.

APPENDIX 1

EXAMPLES OF CONJUGATIONS IN FRENCH SHOWING PRONOUN USAGE

1. Regular verbs

	PARLER *to speak*	FINIR ***to finish***
je	parl**e**	fin**is**
tu	parl**es**	fin**is**
il/elle/on	parl**e**	fin**it**
nous	parl**ons**	finiss**ons**
vous	parl**ez**	finiss**ez**
ils/elles	parl**ent**	finiss**ent**

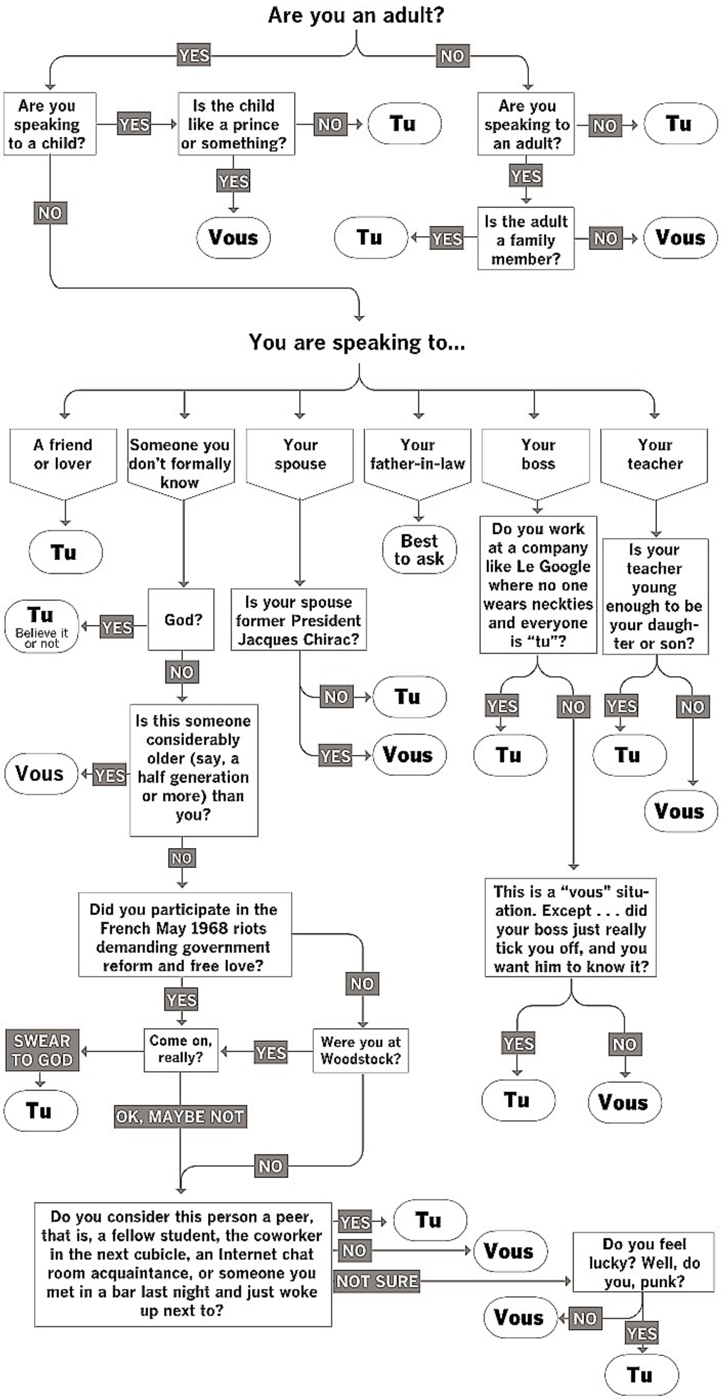
2. Irregular verbs

	ÊTRE *to be*	**AVOIR** *to have*
je	suis	(j') ai
tu	es	as
il/elle/on	est	a
nous	sommes	avons
vous	êtes	avez
ils/elles	sont	ont
	p. 29	*p. 72*

APPENDIX 2

TU OR VOUS USAGE IN FRENCH?

Below you will find guidelines for non-native speakers that teach you when to use *tu* or *vous* in French. Even though this image is somewhat satirical and comes from *BuzzFeed*, it is probably the best and most accurate visualization I have ever seen as a native speaker.


Are you an adult?
YES
NO
Are you speaking to a child?
YES
Is the child like a prince or something?
NO
Tu
YES
Vous
NO
Are you speaking to an adult?
NO
Tu
YES
Is the adult a family member?
YES
Tu
NO
Vous
You are speaking to...
A friend or lover
Tu
Someone you don't formally know
God?
YES
Tu Believe it or not
NO
Is this someone considerably older (say, a half generation or more) than you?
YES
Vous
NO
Did you participate in the French May 1968 riots demanding government reform and free love?
YES
Come on, really?
SWEAR TO GOD
Tu
OK, MAYBE NOT
NO
Were you at Woodstock?
YES
NO
Do you consider this person a peer, that is, a fellow student, the coworker in the next cubicle, an Internet chat room acquaintance, or someone you met in a bar last night and just woke up next to?
YES
Tu
NO
Vous
NOT SURE
Do you feel lucky? Well, do you, punk?
NO
Vous
YES
Tu
Your spouse
Is your spouse former President Jacques Chirac?
NO
Tu
YES
Vous
Your father-in-law
Best to ask
Your boss
Do you work at a company like Le Google where no one wears neckties and everyone is "tu"?
YES
Tu
NO
This is a "vous" situation. Except . . . did your boss just really tick you off, and you want him to know it?
YES
Tu
NO
Vous
Your teacher
Is your teacher young enough to be your daughter or son?
YES
Tu
NO
Vous