

Utilisation d'une méthode hybride pour la  
détection automatique de la polarité dans les  
tweets concernant EDF

Mathilde POULAIN

2017

## Résumé

Ce mémoire traite de la détection automatique de la polarité dans un corpus de tweets traitant d'EDF. Le corpus est constitué des tweets extraits de Twitter via une application développée au sein d'EDF Commerce, appelée TweetsTracker. Cette application réalise plusieurs pré-traitements notamment un filtrage des tweets traitant d'EDF et une classification catégorie par thématique. Ces tweets sont ensuite stockés dans une base de données interne à EDF. Le corpus de l'étude sera composé d'un échantillon de ces tweets respectant la répartition par catégorie.

La méthode consiste à réaliser un lexique de mots de polarité et à construire des règles qui permettront d'extraire des relations en fonction de la présence ou non de mots de polarité. Ces informations sont extraites sur chaque tweets et sont utilisées comme feature pour l'entraînement d'un modèle d'apprentissage automatique.

L'étude du corpus a révélé une corrélation entre certaines thématiques et la polarité des tweets. Nous concluons sur la possibilité de lier thématique et polarité.

**Mots-clés** méthode hybride, Twitter, détection de sentiment, méthode symbolique, apprentissage automatique

**Remerciements** Je tiens à remercier tout particulièrement M. Mathieu VALLETTE pour avoir su gérer avec toute sa sagesse et sa gentillesse mes questions, mes incompréhensions, mes objections et pour m'avoir accompagnée tout au long de ce travail.

J'adresse mes plus sincères remerciements à Anne-Laure GUENET pour m'avoir acceptée pour ce stage, pour la confiance, pour l'accompagnement dans ce travail et pour sa relecture et ses corrections.

Un très grand merci à Sofiane KERROUA qui m'a encadrée durant ce stage avec une patience à toute épreuve, qui a pris tout le temps nécessaire pour répondre à mes questions, débloquer des situations, m'aider à chercher des réponses à chaque fois que j'en avais besoin.

Je souhaite aussi adresser toute ma reconnaissance à toute l'équipe enseignante du master TAL pour ces deux années d'étude, pour tous les cours et conseils reçus qui m'ont permis de découvrir avec émerveillement cette discipline.

Enfin, merci à tous mes collègues d'EDF Commerce, aux stagiaires et alternants de la Direction Numérique et à l'équipe TextMining qui ont participé de près ou de loin à faire de ce stage une aventure enrichissante, pleine de découvertes et d'échanges humains. Merci d'avoir partagé avec moi quelques mois au sein de la Direction Numérique et d'avoir nourris de rires et de débats les repas et pauses durant mon stage.

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Présentation et contexte . . . . .	3
1.2	Définitions . . . . .	4
<b>2</b>	<b>Etat de l'art</b>	<b>5</b>
2.1	Analyse de sentiment et fouille d'opinion . . . . .	5
2.1.1	Présentation . . . . .	5
2.1.2	Quintuple de l'opinion . . . . .	5
2.1.3	Description des méthodes existantes . . . . .	6
2.2	Mise en oeuvre des méthodes existantes . . . . .	7
2.2.1	Prétraitements sur les corpus . . . . .	7
2.2.2	Utilisation/Implémentation d'un lexique . . . . .	8
2.3	Les types de méthodes . . . . .	8
2.3.1	Machine Learning . . . . .	8
2.3.2	Utilisation de règles . . . . .	9
2.3.3	Méthode hybride . . . . .	10
<b>3</b>	<b>Constitution du corpus</b>	<b>13</b>
3.1	Extraction et annotation . . . . .	13
3.2	Répartition du corpus . . . . .	14
3.3	Campagne d'annotation . . . . .	15
3.3.1	Consignes d'annotation . . . . .	15
3.3.2	Pré-annotation . . . . .	16
3.3.3	Résultats de l'annotation et détails du corpus . . . . .	17
3.4	Insuffisance des corpus . . . . .	18
<b>4</b>	<b>Outils et Lexique</b>	<b>20</b>
4.1	XIP : construction de règles symboliques . . . . .	20
4.1.1	XIP et Python . . . . .	21
4.1.2	Choix du lexique . . . . .	22
<b>5</b>	<b>Analyse et mise en oeuvre</b>	<b>24</b>
5.1	Analyse exploratoire . . . . .	24
5.1.1	Tweets émanant d'EDF . . . . .	24
5.1.2	Vocabulaire spécifique . . . . .	24
5.1.3	Particularités par catégorie . . . . .	25
5.1.4	Les tweets d'informations . . . . .	28
5.1.5	Conclusion . . . . .	29
5.2	Construction des règles . . . . .	30
5.3	Traits supplémentaires . . . . .	31
<b>6</b>	<b>Résultats</b>	<b>32</b>
6.1	Baseline . . . . .	32
6.2	Apport de XIP . . . . .	32
6.3	Réseaux de neurones . . . . .	34
6.3.1	Modèle négatif/positif . . . . .	34
6.3.2	Modèle négatif/neutre . . . . .	34
6.3.3	Modèle positif/neutre . . . . .	34

<b>7 Conclusion</b>	<b>35</b>
7.1 Croisement des résultats pour l'intégration dans TweetsTracker .	35
7.2 Des thématiques intrinsèquement liées à une polarité . . . . .	36
7.3 Evolution des thématiques sur Twitter . . . . .	36
<b>8 Annexes</b>	<b>39</b>
<b>A Scripts d'extraction des tweets</b>	<b>39</b>
<b>B Guide d'annotation</b>	<b>44</b>
<b>C Extrait du corpus annoté</b>	<b>57</b>
<b>D Liste des traits et relation renvoyés par XIP</b>	<b>62</b>
<b>E Termes les plus spécifiques</b>	<b>65</b>
E.1 Corpus positif . . . . .	65
E.2 Corpus négatif . . . . .	65
<b>F Résultats détaillés</b>	<b>66</b>
F.1 Régression logistique . . . . .	66
F.2 Bayes Multinomial . . . . .	66
F.3 Bayes Gaussien . . . . .	66
F.4 K plus proches voisins (n=3) . . . . .	66
F.5 Arbres de décision . . . . .	67
F.6 SVM . . . . .	67

# 1 Introduction

## 1.1 Présentation et contexte

Ce mémoire traite de la détection automatique de la polarité de tweets. Il a été réalisé dans le cadre d'un stage au sein de la Direction Numérique d'EDF Commerce. Ce stage s'inscrit dans une volonté d'étudier l'expression spontanée issue des réseaux sociaux. Le volume des données numériques textuelles, disponibles sur Internet (forums, twitter etc.) augmente chaque année. L'analyse de ces informations, structurées ou non, est, aujourd'hui, un impératif stratégique pour une entreprise telle qu'EDF.

L'analyse de sentiment est une application du Traitement Automatique du Language Naturel (TALN) qui se développe et représente un moyen intéressant pour les entreprises de se tenir informées de leur image sur internet, de leur e-réputation ou encore de se renseigner sur les tendances des internautes, les applications sont donc multiples. Des travaux précédents ont permis l'évolution des méthodes et des outils.

Le point de départ de ce projet est l'application TweetsTracker, développée au sein de la Direction Numérique, qui permet un suivi d'EDF sur Twitter. L'objectif est de mettre en place un modèle de machine learning couplé, ou non, à des règles sémantiques pour détecter la polarité d'un tweet issu de l'application.

Après une introduction générale et la présentation du contexte, ayant eu des conséquences sur le choix des outils et sur l'annotation, nous ferons un résumé dans une première partie, chapitre II, des précédentes études sur la détection d'opinion.

L'apprentissage automatique nécessitant un corpus d'apprentissage, une campagne d'annotation a été réalisée au sein de l'entreprise. Le guide d'annotation mis au point durant le stage, est disponible en annexe. Les détails concernant la constitution du corpus et le processus d'annotation sont explicités dans le chapitre III.

Le "contexte EDF" a aussi un impact sur les outils utilisés pour réaliser cette étude. EDF ayant acquis le logiciel XIP de XEROX, c'est cet outil qui sera utilisé pour réaliser les analyses morphosyntaxiques. Les scripts sont réalisés en python, son utilisation découle aussi du choix d'EDF d'utiliser ce langage. Enfin, l'utilisation de la librairie sci-kit learn repose sur la réputation et la praticité de cette librairie pour des problématiques de TAL. L'ensemble des outils utilisés sont détaillés dans le chapitre IV.

La construction des modèles a nécessité une analyse préalable du corpus et certaines parties du corpus ont été étudiées en détails. L'objectif était de détecter la présence de caractéristiques spécifiques dans les tweets selon leur polarité. L'idée étant in fine d'en déduire des règles pour la détection de l'opinion. L'analyse du corpus ainsi que les règles qui constituent les modèles sont détaillés dans le chapitre V.

Après une présentation des résultats obtenus, dans le chapitre VI, nous nous attarderons sur les conclusions à tirer de cette étude. L'analyse du corpus, la construction des règles ainsi que les résultats offerts par l'apprentissage automatique ont montré une certaine corrélation entre les thématiques abordées dans les tweets et une polarité en particulier. Nous avons cherché à déterminer les types de tweets ayant posés des problèmes lors de l'analyse et leurs particularités. Nous nous sommes enfin posé la question de l'expression d'une opinion

publique sur Twitter, tout cela est exposé dans le chapitre VII.

## 1.2 Définitions

Twitter est un réseau social de micro-blogging qui permet l'envoi gratuit de messages courts appelés tweets sur internet.

Les tweets sont des messages échangés sur Twitter, par les utilisateurs de Twitter, les tweetos. Ils peuvent être retweetés, c'est-à-dire retransmis directement sans modification, et/ou likés autrement dit marqués par un utilisateur. Il est également possible de marquer les tweets au niveau du contenu à l'aide des #hashtags, précédés d'un #, ces contenus marqués peuvent être suivis parmi les trending mondiaux ou nationaux. Cela permet aussi d'indexer les tweets en fonction des hashtags utilisés. Les tweetos peuvent ainsi faire des recherches sur les hashtags et en créer librement. Les trending mondiaux sont une indication des thèmes les plus abordés dans les tweets et donc les sujets brûlants sur Twitter.

Les tweets se caractérisent linguistiquement par :

- des textes courts, de 140 caractères maximum
- l'utilisation d'abréviations génériques ou propres à internet
- un style souvent familier
- la présence d'émoticônes
- la présence de mots-clés spécifiques : hashtags, spécifiés par #
- des mentions indiquées par @ pour s'adresser à d'autres twittos, les citer ou les interpeller.

## 2 Etat de l'art

### 2.1 Analyse de sentiment et fouille d'opinion

#### 2.1.1 Présentation

Audrey Lohard et Dominique Boullier [Audrey and Boullier, 2012] présentent dans leur livre l'analyse de sentiment et font un résumé des méthodes existantes pour traiter la détection de polarité. La première partie de leur présentation tente de répondre aux questions sur ce qu'est la subjectivité, les frontières avec l'objectivité et de définir ce qu'est une opinion, nous ajouterons des interrogations sur la polarité.

La détermination de la subjectivité est définie comme la "détection des indices de la présence du subjectif". Mais cette distinction est aujourd'hui basée dans beaucoup d'études sur une opposition subjectif/objectif utilisée pour une classification des textes en amont de toute détermination de la polarité. Avant toute analyse, nous allons donc repérer les éléments marquants la subjectivité dans un texte et le classer comme "subjectif", à l'opposé des textes n'ayant aucun des marqueurs qui seront classés comme "objectif". Mais la frontière peut être floue entre subjectivité et objectivité, par exemple une opinion "peut être implicite ou avoir des effets négatifs sur une appréciation".

1. Mon téléviseur X est tombé en panne hier.

Dans cette phrase à priori objective, le fait que le téléviseur soit tombé en panne apporte une connotation négative, serait-il de mauvaise qualité? Cela peut devenir une critique de la marque X.

Nous pouvons aussi poser la question de la différenciation entre opinion et évaluation. Une évaluation, bien qu'elle mette parfois en avant les aspects positifs et négatifs de ce qui est évalué, est censée être "objective". Un journaliste peut par exemple rapporter un fait considéré comme "négatif" tout en faisant un rapport objectif. Un critique de film/séries/restaurant/hôtel fait-il une critique objective?

#### 2.1.2 Quintuple de l'opinion

La définition d'une opinion a été établie par Bing Liu, de l'Université of Illinois à Chicago sous la forme d'un quintuple :

- Opinion (o, f, oo, h, t) avec
  - o : object, objet
  - f : feature de l'objet
  - oo : opinion orientation, valeur positive, neutre ou négative de l'opinion ou score
  - h : émetteur de l'opinion
  - t : time, date d'émission de l'opinion

Enfin, cette opinion doit être classée en "positif" ou "négatif". Mais quelques questions apparaissent, à commencer par la question de la neutralité, différente de l'objectivité. Il est possible par exemple comparer les deux exemples suivants qui seront détaillés dans la partie III.1.B :

2. Une panne électrique dans un datacenter de British Airways provoque le chaos"

### 3. British Airways : panne de courant ou externalisation mal maîtrisée ?

Il s'agit ici d'un texte subjectif dans lequel l'opinion détaillée est neutre, ni positive, ni négative. Les détails de la différenciation entre une opinion neutre et mixte (contenant des éléments positifs et négatifs) donnent aussi naissance à de nouvelles catégories plus fines. Beaucoup d'études ne prennent pas en compte ces catégories ou les utilisent comme classe pour les cas où aucune autre opinion n'a pu être détectée. Certains ajoutent des classes supplémentaires et prennent en compte l'intensité de l'opinion pour faire la différence entre "Ce film était bien." et "Ce film était trop bien!". La polarité est donc aussi sujette à discussion.

#### 2.1.3 Description des méthodes existantes

Nous allons ici donner une description rapide des approches possibles avant d'entrer dans le détail des travaux dans lesquels elles ont été utilisées. Parmi les méthodes existantes pour la classification de textes en polarité, trois ressortent : l'approche par règles, l'approche par apprentissage et des méthodes hybrides.

L'approche par règles repose d'abord sur une analyse syntaxique du texte et sur la création d'une grammaire et de règles qui vont décrire les constructions spécifiques à extraire [Maurel et al., 2017]. Le principal problème des règles est qu'elles ne peuvent pas être exhaustives, la langue est vivante, d'une part, elle évolue, et les possibilités de combinaisons de mots sont infinies, d'autre part. De plus, il faut prendre en compte dans l'écriture des règles, les erreurs rencontrées particulièrement lorsque nous nous trouvons face à des corpus d'utilisateurs et non des corpus "professionnels". Les règles sont ainsi souvent soit trop générales donc pas assez précises soit trop précises donc pas assez couvrantes et trop discriminantes [Cori, 2008].

Dans l'approche par apprentissage, la fréquence ou la présence d'un terme, les catégories morphosyntaxiques, la syntaxe ou encore la présence de négation sont les traits les plus utilisés pour l'entraînement d'un algorithme [Bo and Lillian, 2008]. Ainsi, cette méthode consiste à extraire des informations qui seront ensuite utilisées comme traits dans un algorithme pour apprendre à déterminer les plus discriminants pour les textes. Cette phase d'entraînement permet de donner à l'algorithme un texte qu'il pourra, à partir de ce qu'il a appris, considérer comme étant subjectif/objectif et/ou positif/négatif. Cette méthode dépend grandement de la qualité du corpus utilisé pour l'apprentissage et des traits choisis pour l'entraînement. Elle nécessite un corpus annoté (donc un corpus pour lequel on sait déjà s'il est objectif/subjectif, etc) ce qui n'est pas évident et requiert souvent une annotation manuelle qui peut s'avérer coûteuse.

Les méthodes hybrides [Brun et al., 2014], vers lesquelles nous nous dirigerons, sont encore assez peu utilisées mais prometteuses. Contrairement aux méthodes actuelles qui reposent sur des phénomènes linguistiques de surface (étiquetage morphosyntaxiques, informations de surface, etc), le principe est de créer des règles pour repérer des constructions ou des liens syntaxiques au sein d'une phrase et ainsi pouvoir en retirer des informations qui seront utilisées pour un apprentissage automatique. Ces informations auront l'avantage de relever d'une analyse en profondeur de la phrase. Les règles peuvent reposer sur des traits implémentés grâce à un lexique, spécifique ou non au domaine traité.

## 2.2 Mise en oeuvre des méthodes existantes

Les trois méthodes vues précédemment ont en commun la nécessité d'effectuer des prétraitements sur les corpus, ceux-ci peuvent néanmoins être variés et nombreux selon la provenance des textes. Les messages obtenus de Twitter, peuvent être mal rédigés et contenir nombre d'éléments spécifiques au format (hashtags, marque de retweets, etc), ils demandent donc souvent des prétraitements plus ou moins lourds.

Un autre point commun est l'utilisation de lexiques, presque tous les travaux sur l'analyse de sentiment ou la fouille d'opinion reposent en partie sur l'utilisation d'un lexique de polarité, nous le verrons et nous explorerons les lexiques utilisés.

### 2.2.1 Prétraitements sur les corpus

Comme nous l'avons dit, les prétraitements effectués sur les tweets sont déterminants, par exemple la suppression des émoticônes [Go et al., 2009], remplacent les noms utilisateurs et les URLs et suppriment les doubles lettres (par exemple dans "cooooooool"). En effet, ils constatent que l'apprentissage se fait principalement sur les émoticônes et donc que ceux-ci faussent les résultats en ayant un poids trop important. Dans le même esprit, les retweets et les doublons sont supprimés, doublons venant de l'API Twitter qui renvoyait plusieurs fois le même tweet sur une requête. Ils suppriment aussi les tweets avec des indicateurs positifs et négatifs :

4. Target Orientation :( but it's my birthday! :)

Parmi les prétraitements particuliers concernant les corpus de tweets, les travaux de [Brun and Roux, 2014] portent sur la décomposition des hashtags et la détection de leur polarité pour améliorer la classification des tweets. Les hashtags sont considérés comme des "métadonnées, des annotations libres" pour lier un tweet à un domaine particulier. Ils sont aussi souvent la clef pour déterminer l'ironie ou l'humour. Ici, l'objectif est d'extraire la polarité sur les différents aspects du message, les différentes cibles de l'opinion exprimée.

Le corpus, issu d'ImagiWeb, est composé de tweets extraits automatiquement pour étudier l'image de personnalités politiques ou de compagnies, il a été annoté manuellement en polarité selon des catégories (apparence, projet politique, communication ...), le corpus est composé de 20000 tweets selon 10 catégories et 17 sous-cibles.

L'extraction des hashtags est simple puisqu'ils sont repérés par #, on peut les classer selon 4 catégories :

- un mot de la langue (analysable avec un lexique)
- un nom propre (analysable avec un lexique ou de la reconnaissance d'entités nommées)
- un groupe de mots (par exemple dans #JeSuisCharlie)
- les impossibles à décoder, acronymes ou chiffres

Plusieurs méthodes permettent de gérer la segmentation selon la catégorie du hashtag, il est possible, par exemple, de repérer les majuscules ou la sous-chaine la plus longue appartenant à un lexique. Une polarité est attribuée aux hashtags et ils seront ensuite intégrés à un lexique. Ce lexique deviendra un élément de la grammaire d'opinion. Leur évaluation indique qu'ils améliorent l'exactitude d'environ 2% et le taux d'amélioration passe à 4% en se concentrant sur les

tweets contenant des hashtags. Cela montre l'importance et la pertinence de ce prétraitement.

### 2.2.2 Utilisation/Implémentation d'un lexique

Presque toutes les méthodes reposent sur l'utilisation de lexiques, il en existe plusieurs : General Inquirer, Sentiwordnet, Opinion finder, NTU Sentiment dictionary pour l'anglais, nous y reviendrons dans une prochaine partie. Il peut également s'avérer intéressant d'utiliser des dictionnaires construits ad hoc à partir d'un corpus. L'utilisation d'un lexique seul se heurte à plusieurs limites :

- non prise en compte du contexte
- non traitement de la négation
- non traitement des figures rhétoriques (ironie, sarcasme ...)

Pour la campagne DEFT<sup>1</sup> [Hamon et al., 2015], les lexiques utilisés étaient nombreux (ANEX, LIDILEM, CASOAR, EMOTAIX, FEEL, Polarimots, DES<sup>2</sup>). Les systèmes présentant les meilleurs résultats étaient ceux utilisant des SVM avec des patrons morphosyntaxiques, les caractéristiques des tweets et le lexique FEEL.

Certains travaux se sont basés sur l'utilisation du lexique Twittartr qui propose une liste publique de mots-clés (174 mots positifs et 185 mots négatifs, n'est plus disponible). Ceci a permis la mise au point d'une baseline reposant sur l'utilisation d'un lexique [Go et al., 2009]. La méthode consiste à compter pour chaque tweet le nombre de mots positifs, le nombre de mots négatifs et à calculer un score de positivité et de négativité, en cas d'égalité, le tweet est annoté positivement par défaut.

## 2.3 Les types de méthodes

### 2.3.1 Machine Learning

De nombreuses études ont été faites en utilisant du Machine Learning, nous présentons ici quelques articles en commençant par les corpus et les prétraitements effectués puis les résultats ou conclusions obtenus à l'issue des études.

La collecte du corpus a bien évidemment une influence sur les résultats obtenus. Dans les travaux de [Pak and Paroubek, 2011], le corpus a été collecté sur Twitter et classifié selon la présence d'émoticônes :, ;, ;) :( ;,(. A cela a été ajouté un ensemble de messages classés comme objectif à partir de tweets de journaux (New York times, Washington Post). Sur ce corpus a été entraîné un classifieur, on considère ici que l'émotion exprimée par l'émoticône est celle du message dans sa globalité. A l'inverse, d'autres [Go et al., 2009] font le choix de ne pas inclure les tweets neutres ou sans opinions. Après prétraitements, ils obtiennent un corpus final de : 800 000 tweets positifs et 800 000 négatifs et font une annotation manuelle de 177 tweets négatifs et 182 positifs.

Pour l'entraînement des algorithmes, l'importance repose sur le choix des traits que celui-ci doit prendre en compte pour l'apprentissage. [Go et al., 2009] utilisent pour l'entraînement d'un classifieur bayésien naïf, à entropie maximal et d'un SVM les traits suivant :

- les unigrammes

---

1. Défi Fouille de Texte

2. Dictionnaire Electronique des Synonymes

- les bigrammes
- une combinaison des unigrammes et bigrammes
- les unigrammes combinés aux POS

Leurs travaux ont donné naissance à un site<sup>3</sup> permettant de visualiser les tweets positif et négatifs à partir de la recherche d'un terme en temps réel. Leur méthode repose sur des classifieurs bayésiens, à Maximum d'Entropie et un SVM. Leur baseline correspond à l'utilisation de Twittratr qui attribue un score à chaque texte.

Plus largement, parmi les équipes ayant participé au DEFT 2015, [Hamon et al., 2015], les traits utilisés étaient les suivants :

- critères stylométriques
- patrons syntaxiques
- descripteurs sémantiques relevés par une analyse textométrique préalable
- présence de la négation
- liste de termes de polarités
- algorithme de clustering non supervisé pour créer des vecteurs de mots notamment avec word2vec
- caractéristiques des tweets : émoticônes, hashtags, nom d'utilisateurs ...

Enfin, les résultats permettent de mettre en avant le classifieur le plus pertinent et les traits qui se sont révélés fondamentaux ou ceux qui ont, au contraire, altéré les performances de certains systèmes. Dans [Pak and Paroubek, 2011], c'est le classifieur Naïve Bayes qui a donné les meilleurs résultats comparés au SVM et au CRF. L'évaluation a été faite en utilisant les mesures d'entropie et de "saillance" (saliency en anglais). Leurs résultats montrent que l'utilisation de bigrammes fonctionne mieux que la combinaison des unigrammes et des trigrammes et que l'élargissement du corpus améliore les résultats.

De leur côté, [Go et al., 2009] constatent que l'utilisation des bigrammes, utilisés pour permettre de prendre en compte, entre autre, la négation, faisait baisser les résultats et que les POS s'avéraient aussi inutiles. Ils proposent comme perspectives l'utilisation de la sémantique pour déterminer le sujet des verbes, une spécialisation limitée à un domaine, la prise en compte des émoticônes et enfin une prise en compte des tweets neutres.

### 2.3.2 Utilisation de règles

Dans les travaux de [Brun, 2011], l'accent est mis sur la nécessité d'effectuer des extractions précises plutôt que d'accorder une polarité globale à un message ou un texte. On parle de concepts principaux et de concepts associés. Il s'agit alors d'extraire une polarité par rapport à un élément précis lié au sujet général du message.

Lors de ce travail sur 3 500 avis consommateurs sur des imprimantes extraits du site [epinions1](http://www.epinions1), l'analyse exploratoire montrait que les phénomènes linguistiques mis en jeu sont la présence de la négation, l'importance de la modalité et l'utilisation de comparaisons. Elle met aussi en avant l'ambiguïté d'un lexique de polarité selon son contexte :

5. It uses ink twice as fast. [opinion négative]
6. It is also a fast, high quality printer. [opinion positive]

---

3. <http://www.sentiment140.com/>

L'extraction de concepts principaux et des concepts associés nécessite la mise en place de tâches spécifiques : l'extraction des concepts, la découverte des opinions et la gestion de toutes ces opinions pour fournir potentiellement une opinion globale. En l'occurrence, nous construisons une architecture à deux niveaux avec la création d'un vocabulaire général de polarité et un spécifique au domaine étudié.

C'est l'outil XIP (Xerox Incremental Parser) qui est utilisé, il permet notamment une analyse syntaxique profonde et la création de règles pour l'extraction de relation de dépendances. Dans celui-ci ont donc été implémentées des règles pour extraire des relations sémantiques du type :

```
SENTIMENT [POLARITE](concept_principal, concept_associe,
    predicat_porteur)
```

Par exemple :

7. The printer is slow → SENTIMENT [NEGATIF] (printer, -, slow)

Le concept\_associe est nul car l'opinion exprimée porte sur l'imprimante dans son ensemble. Dans :

8. The laser print quality is great. → SENTIMENT [positif] (default\_product, print quality, great)

Le concept\_principal n'est pas indiqué donc par défaut ce sera le produit sur lequel porte la revue, le concept associé est indiqué et le prédicat est "great".

L'extraction des concepts associés repose sur la recherche de relations de méronymie, de noms modifiant un concept principal selon une relation de modifieur antéposé et de modifieur lié par la préposition "of" entre autres.

```
If(SUBJ-N(#1[polarite!polarite:!,topic-subj],#2[concept-principal]))
    SENTIMENT(#1, -, #2)
```

Dans cette règle, on affecte une relation de sentiment aux verbes de polarité et à son sujet si c'est un concept principal.

Exemple reconnu par cette règle et relation créée :

9. These printers never cease to amaze me. → SENTIMENT [positif] (printer, -, amaze)

Une soixantaine de règles ont été développées. A cela, il faut ajouter des règles de traitement de la négation qui inversent la relation de sentiment existante, une vingtaine de règles spécifiques au domaine des imprimantes et une dizaine de règles sur la structure spécifique des commentaires d'opinions.

L'évaluation montre que le système présente plus de difficultés dans la catégorisation des revues négatives mais les résultats sont satisfaisants. Se pose néanmoins la question d'un potentiel problème de couverture du vocabulaire négatif. Le constat montre que dans les avis défavorables, les utilisateurs ont tendance à utiliser la comparaison avec un produit qu'ils ont apprécié d'où l'importance d'une implémentation de règles pour la reconnaissance des comparaisons.

### 2.3.3 Méthode hybride

Dans la continuité de [Brun, 2011], un exemple de méthode hybride est donné dans [Brun et al., 2014]. L'objectif de déterminer un sentiment associé

aux différents aspects d'un même objet et donc de faire une analyse beaucoup plus fine des sentiments demande ici de s'attaquer à la détection des cibles de l'opinion exprimée.

Ce travail sur un corpus anglais de commentaires d'internautes fournis par SemEval contient comme prétraitement une suppression des mots-outils et se divise en 4 tâches :

- Détection des aspects
  - I loved their fajitas → fajitas
- Extraction des catégories sémantiques dont relèvent les aspects trouvés
  - I loved their fajitas → fajitas : food
- Extraction de la polarité des aspects (positif, négatif, neutre, conflit)
  - I loved their fajitas → fajitas : positif
- Extraction de la polarité des catégories
  - I loved their fajitas → food : positif

Suite à ce qui est détaillé ici en II.3, [Brun, 2011], basé sur XIP, qui permettait l'extraction de dépendances syntaxiques et des informations lexicales pour générer des relations d'opinion, ils obtenaient des dépendances sémantiques appelées SENTIMENT de la forme :

SENTIMENT[NEGATIF](disappointed, service)

Les adaptations faites au système incluent une adaptation du lexique avec la création d'un lexique spécifique au domaine traité, une adaptation de la grammaire, modifiée pour pouvoir déterminer quand un terme est la cible de deux relations SENTIMENT et si ces deux relations ont la même polarité.

Une première étape permet la détection des termes et de leur polarité, pour laquelle est utilisée une grammaire, puis une deuxième étape de détection des catégories et de leur polarité pour laquelle est utilisée cette fois une classification automatique.

Pour cette deuxième étape, un premier classifieur associe une catégorie sémantique aux phrases puis un deuxième classifieur travaille sur chaque catégorie pour leur associer une polarité. L'un fonctionne avec des sacs de mots et les informations fournies par XIP, l'autre fonctionne sur un modèle de régression logistique. La polarité neutre est automatiquement associée au texte si aucune des autres n'y a été associée.

Une dernière étape permet de revenir sur la polarité des termes. L'objectif est de corriger le fait que la polarité neutre soit affectée si la catégorie sémantique n'a pas été détectée comme positive ou négative. Cette correction s'est révélée très intéressante pour les résultats.

La baseline utilisée pour l'évaluation consistait à utiliser un lexique de terme construit ad hoc à partir du corpus d'entraînement et à les repérer dans le corpus de test pour calculer un score de polarité. La détection des catégories se fait par un calcul des  $k$  plus proches phrases du corpus d'entraînement, elles sont ensuite associées aux phrases du corpus de test.

Leurs résultats sont nettement meilleurs pour la détection des catégories et des termes, de même pour la tâche de détection de la polarité avec une exactitude globale de 0,78 pour la détection de la polarité des termes et des catégories. Néanmoins, ce système peine à détecter la polarité conflit et, la polarité neutre étant affectée par défaut fait, semble-t-il, baisser les résultats. Plusieurs expériences ultérieures ont montré l'intérêt de la prise en compte des  $n$ -gram et de la distance entre un terme ou une catégorie et les termes de polarité.

Une deuxième méthode hybride utilisant des arbres de dépendances a été expérimentée dans [Pak and Paroubek, 2011]. Elle repose aussi sur XIP et veut utiliser les relations de dépendances entre les mots, cette information étant perdue lors d’une approche à base de sac de mots alors que les relations syntaxiques permettraient une meilleure analyse et une prise en compte de la construction d’une phrase pour détecter l’opinion.

La génération des sous-graphes se fait par simplification et réduction des arbres générés par XIP. Certaines relations sont regroupées, d’autres supprimées car inutiles pour l’analyse d’opinion. La première simplification consiste à regrouper les relations dans leur classe générique, ainsi : SUBJ\_COORD, SUBJ\_PASSIVE ou encore SUBJ\_PROPQUE seront rassemblées sous le nom de relation SUBJ. Les relations supprimées sont celles n’ayant pas d’impact sur l’analyse de sentiment par exemple la relation de détermination ou les relations de modifieur entre deux noms (“livre de cuisine”).

L’étude porte sur un corpus de critique de jeux vidéos réunis à partir de 8 sites différents qui ont été annotés entre autres avec les indications d’une catégorie sémantique, d’une polarité (+, +-, - et neutre) accompagné d’une intensité (très négatif, faible-moyen-négatif, mixte, etc) et d’un thème qui précise la cible de l’opinion exprimée. Parmi ce corpus, un sous-corpus contenant des critiques positives et négatives a été construit, les critiques neutres ne sont pas intégrées, et celui-ci a été encore subdivisé pour former un corpus d’entraînement et un corpus de test.

La baseline a été calculée avec un classifieur SVM et un schéma de pondération delta TFIDF et entraîné sur des unigrammes, des bigrammes et des trigrammes pour construire trois modèles. Leur propre modèle a été construit avec des sous-graphes de taille 1, 2 et 3.

La baseline donne une exactitude de 73.86% avec des unigrammes et leur modèle donne les meilleurs résultats avec des sous-graphes de taille 1 pour une exactitude de 78.41%. Le corpus d’entraînement étant un ensemble de 531 documents, il est considéré comme potentiellement insuffisant pour la construction du modèle, de plus il est constitué de 334 documents positifs et de 197 documents négatifs, un déséquilibre qui se répercute sur les résultats puisque ceux-ci sont moins bons pour les documents négatifs.

## 3 Constitution du corpus

### 3.1 Extraction et annotation

Les tweets sont récupérés via l'application TweetsTracker, mise au point par Sofiane KERROUA, et stockés par EDF, nous y ferons référence comme étant le corpus EDF. Depuis sa mise en place, plus de 900 000 tweets ont été récupérés. Pour ce projet, nous avons donc à disposition un corpus conséquent mais l'évaluation des résultats nécessite un corpus annoté, ce qui sera fait manuellement. Cela exclu donc la possibilité d'utiliser l'ensemble des tweets mis à disposition grâce à TweetsTracker, nous avons donc du prendre un échantillon du corpus EDF.

TweetsTracker récupère les tweets à partir de l'API Twitter et réalise plusieurs traitements. D'abord, les tweets sont récupérés par extraction de certains mots-clés. Ensuite, un filtrage est effectué afin de différencier les tweets traitant d'EDF (Electricité de France) et EDF (Equipe de France). Ce filtre mis au point par l'équipe de la R&D est une étape préalable au stockage des tweets dans la base de données. Nous supposons donc que la base de données ne contient que des tweets traitant d'EDF.

Un autre traitement consiste à faire une recherche des insultes afin de classer certains tweets comme "détracteurs". Cette classification repose uniquement sur une recherche de termes. Ce projet veut permettre une détection plus fine des opinions exposées sur Twitter. Cependant, ce traitement offre une première base pour constituer un lexique d'insulte et donc caractériser les tweets détracteurs comme négatifs.

Ce corpus a aussi été annoté en catégories selon le sujet abordé, les catégories sont les suivantes :

- Actualités
- Démarchages frauduleux
- Contrat
- Coupure
- Environnement
- Intervention Technique
- Linky
- Nucléaire
- Paiement facture
- Relation Client
- Tarif
- TRV

Il est possible pour un tweet d'appartenir à plus d'une catégorie et certains messages, ne rentrant dans aucune catégorie, sont non classés et n'ont donc aucune étiquette.

Chaque tweet est récupéré avec les informations suivantes :

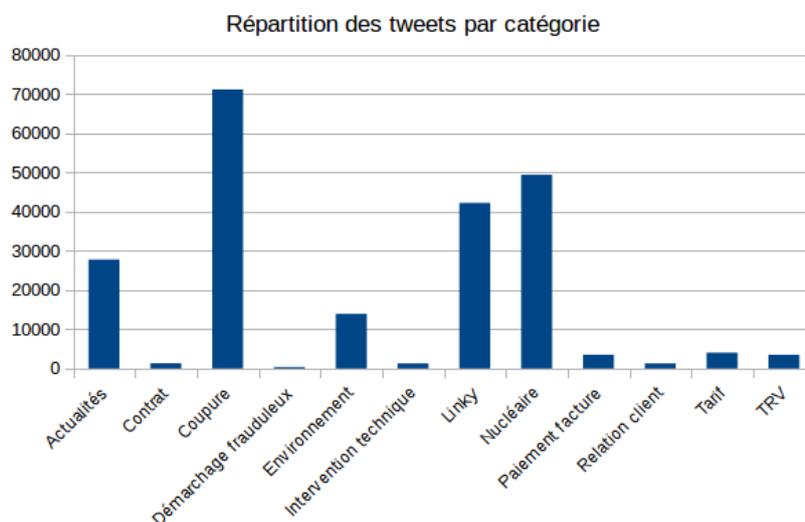
- un indentifiant
- la date de publication
- le nom de l'auteur (ou pseudo)
- les URLs
- le lemme de chaque terme
- le tweet lemmatisé
- l'information d'un retweet

- la catégorie attribuée
- le texte du message

Dans le corpus constitué pour ce projet (le corpus de travail) la répartition des tweets par catégorie est respectée. On reproduit donc dans cet échantillon la même distribution par catégorie que dans le corpus EDF, ceci afin d’avoir un ensemble aussi homogène dans sa globalité. Une première étape de la constitution du corpus est donc de déterminer la répartition des tweets par catégorie.

### 3.2 Répartition du corpus

Lors d’une première analyse, la base de données contient plus de 900 000 tweets. Sur l’ensemble des tweets qui sont contenus entre mai 2016 et mai 2017, la répartition des tweets par catégorie est indiquée de le graphique suivant.



Cette première visualisation du corpus permet d’obtenir une représentation des thèmes les plus récurrents. Cela justifie l’utilité de respecter la répartition dans le corpus de travail afin d’éviter une spécialisation du modèle face à des tweets traitant d’un thème spécifique (et une discrimination trop forte à partir de certains traits).

Il est envisageable qu’une extraction, même aléatoire, retourne principalement des tweets d’une seule et même catégorie. Couplé à une limite due à l’annotation humaine qui est faite, nous ne pourrons pas utiliser un nombre trop élevé de tweet, et au vu de cette distribution, il est probable, mais problématique, d’obtenir un corpus déséquilibré.

Afin d’éviter cet écueil, nous avons mis en place un script d’extraction des tweet stockés dans la base de données selon plusieurs critères. Tout d’abord nous nous concentrerons sur les tweets compris entre mai 2016 et mai 2017 et les tweets seront extraits au nombre de 2250 en respectant la répartition des tweets dans la base de données. Cela afin d’obtenir un corpus ayant le même pourcentage de tweets dans chaque catégorie que le corpus EDF. Le script d’extraction est disponible en annexe.

Le nombre de 2250 est obtenu en partant d’une annotation de 500 tweets par personne plus une personne annotant 1000 tweets. Avec 8 annotateurs, on

obtient un corpus de 2250 tweets avec une double annotation.

### 3.3 Campagne d'annotation

#### 3.3.1 Consignes d'annotation

Dans l'optique de la création d'un modèle, avoir à disposition des données annotées est essentiel pour l'évaluation, la vérification du fonctionnement et l'amélioration du système. Les tweets de TweetsTracker sont annotés en catégorie mais pas en polarité et la classification des tweets détracteurs est insuffisante. Nous avons donc mené une campagne d'annotation au sein de l'équipe TextMining d'EDF pour mettre au point un corpus annoté.

L'annotation sera effectuée par 8 personnes, 7 personnes annoteront 500 tweets chacune et une personne annotera 1000 tweets. Pour les besoins d'une double annotation, les ensembles de 500 tweets sont divisés en deux parties qui seront annotées par deux personnes. Ainsi, pour chaque annotateur, sur l'ensemble de ses 500 tweets à annoter, la première partie sera annotée par une deuxième personne, et la deuxième partie sera annotée par une troisième personne. Nous obtenons un corpus final de 2250 tweets avec une double annotation.

Le guide d'annotation contenant l'ensemble des consignes d'annotation et la définition des catégories ainsi que quelques détails sur les cas particuliers ou complexes est disponible en annexe, nous en faisons ici un bref résumé.

Dans le cadre de cette annotation, nous avons défini de façon générale les différentes annotations possibles et des consignes pour les cas les plus compliqués. Ainsi, au-delà des définitions assez simples et intuitives que l'on peut se faire d'un tweet positif, négatif ou neutre, nous allons nous attarder ici sur les cas particuliers.

Avant toute chose, nous considérons comme négatif un tweet présentant une information négative ou faisant part d'un point de vue négatif. De même, nous considérons comme positif un tweet présentant une information positive ou faisant part d'un point de vue positif.

Notre corpus présente un nombre très élevé de tweets liés à des coupures de courant. Bien qu'ils soient très souvent lié à un mécontentement du à la coupure de courant, nous n'avons pas systématiquement annoté négativement tous les tweets traitant des coupures de courant. Certains tweets seront considérés comme étant informatifs ou tout simplement neutre :

10. Petit coupure de courant
11. Y aurait plus d'électricité entre Paris Nord et la banlieue
12. y'a eue une panne d'électricité je crois..

Nous nous sommes attardés sur ce que nous avons appelé les "tweets d'information". Sont englobés dans ce concept tout message qui aura généralement été posté par des organes d'informations comme les journaux ou magazines d'informations et qui présente une actualité. Cela concerne également les tweets émanant de particuliers s'ils font circuler des liens vers des articles d'actualité ou les messages issus de sites de ré-information. Nous détaillons ici l'annotation pour ces tweets particuliers, souvent à mi-chemin entre de l'information objective et une prise de position par rapport à cette information.

Une information négative ou positive présentée de façon neutre reste une information négative ou positive, elle est donc annotée comme tel :

13. "Manifestations contre Linky, Gazpar ont lieu dans toute la France" est annoté négativement.

Les informations à priori objectives présentées selon un point de vue et avec une prise de position pour ou contre sont annotées selon la prise de position de l'auteur :

14. " Comprenez-vous que l'on débaptise tous nos fleurons nationaux (TGV, EDF etc.) pour y accoler des noms anglais ? " *négatif*

Une information peut donc être *objective ou orientée* et être présentée de *façon objective ou orientée*. Pour résumer et illustrer cette problématique, prendre en considération l'exemple suivant :

15. "Une panne électrique dans un datacenter de British Airways provoque le chaos" *négatif*

versus :

16. "British Airways : panne de courant ou externalisation mal maîtrisée?" *neutre*

Dans le premier cas (15), l'information est présentée de façon négative, les faits ont provoqué le "chaos" et la cause de ce "chaos" est une panne de courant. Par conséquent, le tweet est annoté comme négatif. Dans le second cas (16), l'information est objective (elle ne critique pas ni ne remet en cause EDF mais British Airways) et présentée de façon neutre. Elle informe sur une panne de courant mais pose la question des causes sans poser d'accusations. De plus, le choix a été fait de ne pas systématiquement annoté négativement les tweets concernant les pannes de courant. Pour toutes ces raisons, le tweet est donc annoté comme "neutre". Une même information, un même fait peut donc être annoté de deux façons différentes selon la polarité du message qui l'accompagne.

### 3.3.2 Pré-annotation

Afin de vérifier et, le cas échéant, de compléter le guide d'annotation, une première étape de pré-annotation a été mise en place. Pour ce faire, tous les annotateurs ont annoté les 50 mêmes tweets afin de vérifier l'homogénéité de l'annotation et de contrôler l'accord inter-annotateur.

Voici quelques chiffres sur les résultats obtenus lors de cette première étape, l'accord inter-annotateurs fait ici référence au nombre de réponses communes obtenues par rapport au nombre de réponses communes attendues :

$$P = \frac{\text{nombre de réponses attribuées à la même classe par les annotateurs}}{\text{nombre de réponses appartenant à la classe dans le corpus}} * 100$$

	<b>A1</b>	<b>A2</b>	<b>A3</b>	<b>A4</b>	<b>A5</b>	<b>A6</b>
<b>A1</b>		80%	72%	78%	86%	76%
<b>A2</b>			82%	78%	78%	84%
<b>A3</b>				72%	70%	78%
<b>A4</b>					84%	82%
<b>A5</b>						82%

La pré-annotation montre un peu plus de 50% de réponses communes aux 8 annotateurs et un accord inter annotateurs en moyenne de 78.8%. Nous avons

considéré que c'était suffisant pour lancer une annotation sachant que chaque extrait serait annoté par deux personnes.

Cette pré-annotation a toutefois permis de mettre en lumière la complexité de l'annotation de l'ironie, souvent difficile à détecter. La question a aussi été posée concernant des tweets ayant été récupérés dans le corpus mais pour lesquels le thème n'est pas forcément ni explicitement EDF. Si ces tweets sont clairement négatifs mais qu'ils ne traitent pas d'EDF, une annotation neutre risque de fausser le modèle. En effet, le risque est de se trouver face à des messages ayant toutes les caractéristiques d'un tweet négatif, mais en étant classé comme neutre. Cela peut donc amener à faire baisser le potentiel discriminant des caractéristiques malgré leur véritable utilité.

D'autre part, comme expliqué lors de la constitution du corpus, un système existant filtre les tweets traitant ou non d'EDF. Ainsi notre corpus est déjà supposé ne contenir que des tweets à propos d'EDF. Inclure dans ce modèle sur la polarité un deuxième filtrage des tweets selon leur thème ne nous paraît pas pertinent et demanderait un travail supplémentaire trop important. Celui-ci ne sera donc pas réalisé ici.

Le parti pris final est de les classer en polarité, cette solution semble préférable pour la construction du modèle. En ce qui concerne les tweets ironiques, une relecture finale sur les tweets pour lesquels les annotateurs ont eu des doutes sera faite.

Voici quelques exemples de tweets auxquels ont été attribuées les trois polarités (positif, négatif et neutre) par les annotateurs :

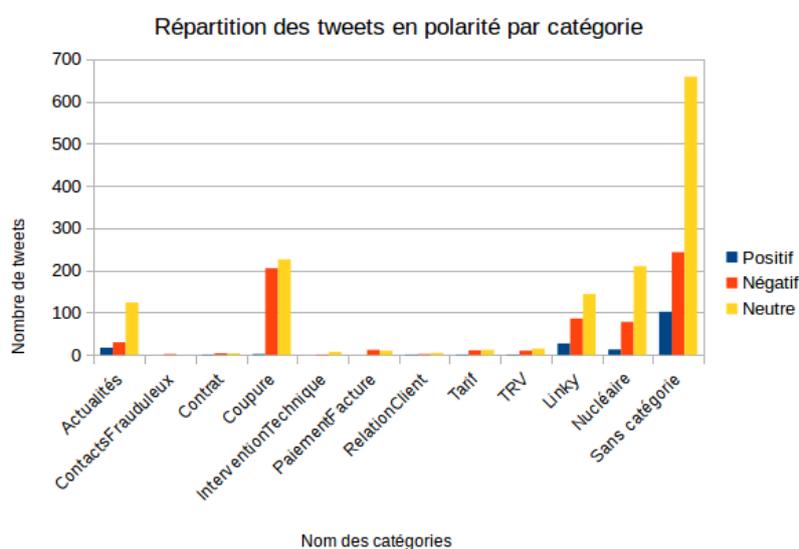
Tweet
Oui coupure de courant @NanzarFr @ReisGreg Attention à vous :confused_face : :thumbs_up :
@KuntaAB j'ai pas d'électricité la :smiling_face_with_smiling_eyes : y a une coupure de courant dans tout l'immeuble et on est le seul bureau chez qui les ordi fonctionnent je vais :
:red_circle :ERDF Demain, ERDF, filiale d'EDF, sera renommée Enedis, à la demande de la Commission de régulation de l'énergie. Coût : 300 M (@LesEchos)
@A0R_AR30OoB @electroallergik pourquoi avoir coupé l'extrait juste avant le passage de la mesure avec CPL? #Linky URL

### 3.3.3 Résultats de l'annotation et détails du corpus

A partir de la répartition dans le corpus EDF et des impératifs d'une annotation manuelle, nous avons extraits 2250 tweets aléatoirement sur une période allant de mai 2016 à mai 2017 en sélectionnant les tweets afin que le corpus final corresponde à la répartition par catégorie.

Le corpus de travail est donc constitué de 2250 tweets dont la répartition en polarité et par catégorie est donné dans les figures suivantes.

Catégorie	Positifs	Négatifs	Neutre	Total
Actualités	17	30	124	171
Contacts Frauduleux	0	2	0	2
Contrat	1	4	4	9
Coupure	2	205	226	433
Intervention Technique	0	1	7	8
Paiement Facture	0	12	10	22
Relation Client	1	2	5	8
Tarif	1	11	12	24
TRV	1	10	15	26
Linky	27	86	144	257
Nucléaire	13	78	210	301
Sans catégorie	102	243	658	1003



Pour cette étude, deux modèles seront constitués, un pour les tweets positifs et un pour les tweets négatifs. Nous avons donc, à partir du corpus constitué précédemment, séparé les corpus afin de pouvoir faire un apprentissage différent sur chaque polarité. Pour le corpus négatif, c'est-à-dire le corpus utilisé pour la construction d'un modèle pour la détection des tweets négatifs, les tweets ayant été annotés positivement sont changés en neutre. Nous obtenons un corpus contenant uniquement des tweets négatifs ou neutres. La même opération a été effectuée sur le corpus positif.

A partir de ces deux corpus, nous avons pu construire deux modèles d'apprentissage automatique utilisant comme features des informations issues de XIP.

### 3.4 Insuffisance des corpus

Les premiers tests ont rapidement montré l'insuffisance des corpus. Sur l'ensemble des 2250 tweets, seuls 172 tweets sont positifs dans notre corpus contre 665 négatifs. Un tel déséquilibre ne permet pas un apprentissage suffisant et

compromet les capacités du modèle.

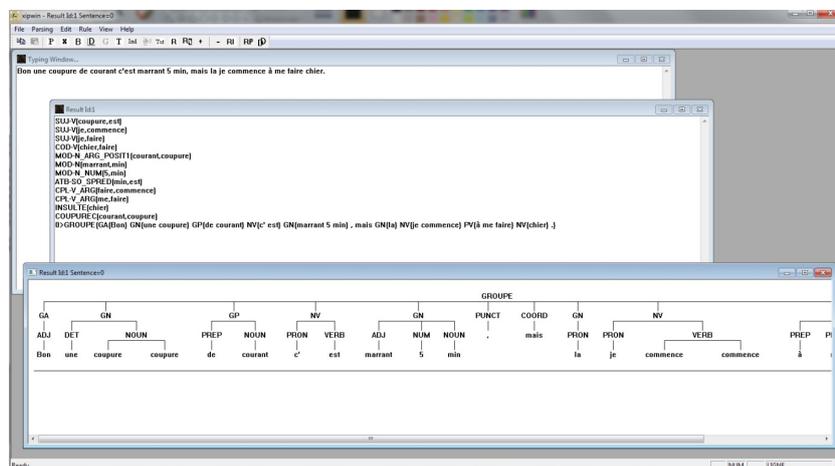
Pour pallier à ce déficit en tweets positifs, une deuxième annotation a eu lieu sur une nouvelle extraction de tweets depuis la base de données. Après une nouvelle séance d'annotation, le nombre de tweets positifs s'élevait à 450. En utilisant tous les autres tweets du corpus comme neutres, on se retrouve tout de même avec un corpus composé de 450 tweets positifs et 1800 tweets neutres pour l'apprentissage, ce qui n'est toujours pas idéal pour un apprentissage.

Pour remédier à ce problème, autant pour les tweets positifs que négatifs, nous avons procédé à un rééquilibrage des corpus. Nous avons donc entraîné le modèle positif sur un ensemble de 900 tweets et le modèle négatif sur un ensemble de 1330 tweets afin d'avoir des corpus composé à 50% de tweets en polarité et de tweets neutres. Malgré un reste de déséquilibre entre les données d'entraînement des deux modèles, ce choix a permis un meilleur apprentissage pour chacun des modèles et donc de meilleures prédictions.

Une fois le corpus finalisé, dont un extrait est présenté en annexe, nous sommes passés à la construction des modèles. Nous avons explicité le souhait de mêler les techniques de machine learning à des approches symboliques pour cette étude. La partie machine learning sera réalisée en python et les règles seront écrites via le logiciel XIP.

## 4 Outils et Lexique

### 4.1 XIP : construction de règles symboliques



L'analyse automatique d'une phrase repose sur un ensemble de règles et les différentes combinaisons de ces règles. Dans XIP (Xerox Incremental Parser), l'application des règles suit un ordre définit.

Le fonctionnement de XIP, qui repose sur des transducteurs à états finis, se rapproche du fonctionnement de la grammaire HPSG, dans la mesure où tout doit être déclaré et tout est déclaré comme une structure de traits. De même, la grammaire HPSG étant une grammaire de tête, XIP détecte les têtes au sein de chaque syntagme.

La première étape du fonctionnement interne de XIP consiste à implémenter tous les traits et toutes les catégories qui seront utilisés lors de l'analyse.

Vient l'étape d'implémentation des lexiques, c'est à ce moment qu'on indique les traits de chaque mot du vocabulaire. Le trait est une information que l'utilisateur ajoute à un terme pour l'enrichir. Un terme peut avoir plusieurs traits. Il peut s'avérer nécessaire d'ajouter un trait "digital" aux termes relevant du champ sémantique du digital, cela s'écrit ainsi dans la syntaxe XIP :

17. site += [digital=+]. (on ajoute le trait digital au terme "site")
18. chien += [noun :+]. (on ajoute le trait "noun" au nom "chien")

Une première grammaire est ensuite appliquée, suivie de règles de désambiguïsation qui vont prendre en compte le contexte des termes afin de lever les ambiguïtés.

Le regroupement en chunk permet la construction des syntagmes, groupes nominaux, verbaux et construit l'arbre syntaxique de la phrase. Les règles de déduction sont appliquées pour créer les relations de dépendances entre les groupes et/ou les termes.

La dernière étape de "renommage des chunks" correspond à l'adaptation de l'analyse réalisée par XIP à la grammaire PASSAGE. On effectue ici un renommage des nœuds et relations syntaxiques de la phrase pour qu'ils correspondent à la dénomination utilisée par la grammaire PASSAGE.

Cette grammaire est le résultat d'un projet ANR en lien avec la campagne d'évaluation EASY du projet EVALDA. C'est un protocole d'évaluation des

analyseurs syntaxiques qui a donné lieu à un processus d’annotation en constituants dont les détails ont été explicités par [?] <sup>4</sup>.

Nous avons développé une grammaire spécifique à partir des traits et principalement des relations renvoyées par la grammaire PASSAGE. L’objectif était de faire ressortir les éléments discriminant pour la détection de la polarité négative ou positive et d’en utiliser les résultats pour entraîner un modèle de machine learning.

Dans notre grammaire, nous travaillons donc au niveau de l’implémentation des lexiques, avec la création de notre lexique en polarité et de nouveaux traits, et au niveau des règles de déduction/dépendances.

Nous implémentons donc un trait de polarité à certains mots dès le lexique et ceux-ci seront ensuite utilisables pour créer et appliquer des règles de déduction. Attribuer un trait de polarité à chaque terme nécessite une déclaration préalable du trait créé, le lexique sera de la forme :

Vocabulary :

bâtard += [negatif=+].  
batifoler += [positif=+].  
battant += [positif=+].  
batteuse += [positif=+].  
bavardage += [negatif=+].

Avec ce trait, il est possible, de créer une règle qui définit un nœud constitué d’un mot ayant le trait “NEGATIF“ :

NOM\_NEG → ?\*, noun[negatif :+].

L’analyse syntaxique profonde du texte, par exemple pour la phrase “Bon une coupure de courant c’est marrant 5 min, mais la je commence à me faire chier.” renvoie les relations suivantes :

SUJ-V(coupure,est)  
SUJ-V(je,commence)  
SUJ-V(je,faire)  
COD-V(chier,faire)  
MOD-N\_ARG\_POSIT1(courant,coupure)  
MOD-N(marrant,min)  
MOD-N\_NUM(5,min)  
ATB-SO\_SPRED(min,est)  
CPL-V\_ARG(faire,commence)  
CPL-V\_ARG(me,faire)  
0; GROUPE{GA{Bon} GN{une coupure} GP{de courant} NV{c’est} GN{marrant 5 min} , mais GN{la} NV{je commence} PV{à me faire} NV{chier}}

#### 4.1.1 XIP et Python

Pour la construction des modèles de machine learning nous avons utilisé Python.

---

4. L’ensemble de la grammaire est présentée ici :  
[https://perso.limsi.fr/anne/Guide/PEAS\\_reference\\_annotations\\_v2.2.html](https://perso.limsi.fr/anne/Guide/PEAS_reference_annotations_v2.2.html)

L'utilisation via python de XIP permet de récupérer tous les traits instanciés dans XIP pour chaque terme, ce qui représente une quantité d'information conséquente.

En reprenant l'exemple précédent, "Bon une coupure de courant c'est marquant 5 min, mais la je commence à me faire chier.", l'ensemble des traits renvoyés dans XIP sont montrés en annexe, en voici un aperçu pour le premier terme :

```
['Bon', 'bon_ADJ', 'bon_[QUEP :+', 'bon_POURSVINF :+',  
'bon_POURSN :+', 'bon_ENVERSSN :+', 'bon_ENSN :+',  
'bon_DESVINF :+', 'bon_DANSSN :+', 'bon_CONTRESN :+',  
'bon_AVECSN :+', 'bon_ASVINF :+', 'bon_ASN :+',  
'bon_!SVINFDIR :+', 'bon_IMPERSO :+', 'bon_STARTBIS :+',  
'bon_MAJ :+', 'bon_SFPOUR :+', 'bon_SFENVERS :+',  
'bon_SFEN :+', 'bon_SFDANS :+', 'bon_SFCONTRE :+',  
'bon_SFAVEC :+', 'bon_SFA :+', 'bon_MASC :+', 'bon_SG :+' ...
```

Nous avons donc d'une part mis au point des règles et implémenté un lexique de polarité en se posant la question de l'intérêt de partir de lexiques existants.

#### 4.1.2 Choix du lexique

Différents lexiques sont aujourd'hui disponibles en français et notre modèle reposera en partie sur les traits de polarité qui seront extraits du lexique. Les lexiques disponibles sont somme toute assez généralistes donc non appropriés au contexte du domaine de l'énergie et en particulier d'EDF. Ils nécessiteront des modifications. Les différents lexiques à disposition c'est-à-dire téléchargeables, utilisables gratuitement et connus sont les suivants :

- Le lexique FEEL contient 14 120 termes, la classification se fait uniquement entre le positif et le négatif. Une classification par catégorie est aussi indiquée : joie, peur, tristesse, etc.
- Le lexique polarimots porte sur 7 483 mots avec l'attribution d'un poids (entre 0 et 100%) pour la polarité positive ou négative et la neutralité.
- Le site internet "jeuxdemots", qui regroupe des jeux dans le but de rassembler des ressources lexicales collaboratives de façon ludique, permet un téléchargement de sa base de données de polarité. Le réseau contient 1 038 425 termes.
- Le lexique SentiWordNet est le résultat d'une traduction automatique du SentiWordNet anglais. Il est au format XML mais ne fournit pas de polarité sur les mots, d'autres formats dérivés du WordNet anglais existent comme WOLF, mais ils s'avèrent être aussi inutilisables.

Le premier choix été de prendre le lexique le plus important dans l'espoir qu'il soit le plus complet. Il semble finalement préférable d'avoir un lexique moins important, comportant donc moins de termes mais qui permettra une meilleure précision dans les termes portant une polarité et étant à la base des relations qui seront extraites. L'utilisation d'un deuxième lexique plus restreint a montré une amélioration des résultats. Nous sommes donc allés plus loin et avons repris entièrement le lexique pour se limiter aux termes porteurs de polarité de façon certaine et vérifiée.

Nous avons donc construit manuellement un dictionnaire ad hoc dont la mise au point sera facilitée par l'utilisation d'un lexique préexistant, en l'occurrence le lexique polarimots. Certains mots seront supprimés notamment des verbes ou

des mots sans rapport avec EDF, ou ne portant pas de polarité dans le contexte d'EDF :

appeler, café, changer, changeur, chanson, chansonnette, vernir,  
manger, achat, fenêtre, ordinateur . . .

Pour certains termes, leur fréquence d'apparition dans le corpus autant dans les tweets négatifs que positifs en font des termes dont la polarité va varier exclusivement en fonction du contexte, il est difficile de leur attribuer une polarité à l'avance, considérons les deux exemples suivants :

19. Après Dampierre, un autre " futoir atomique " à la centrale #nucléaire #EDF de St-AlbanMais ne généralisons pas ... URL
20. tant mieux pour tous les emplois que cette décision préserve,de toute façon le nucléaire reste indispensable,les éc. . . URL

Ces deux exemples montrent que certains termes, en l'occurrence "nucléaire", sont potentiellement autant positifs que négatifs en fonction du contexte, il n'est donc pas possible de leur attribuer une polarité "à priori". Ces termes seront donc supprimés du lexique.

Nous avons ajouté certains mots relevant de ce qu'on pourrait appeler le "vocabulaire Twitter", c'est-à-dire des mots non inclus dans le lexique mais très présents sur Twitter. Nous avons rajouté également les termes relevant d'un niveau de langage exclus des dictionnaires ou des lexiques "classiques" (comprendre non spécifiques à Twitter et au langage des réseaux sociaux) incluant nombre d'abréviations fréquentes dans le corpus et les insultes.

moche, chier, emmerder, énervé, démentis, nsm, ptn, etc

Afin de mettre au point les règles et le lexique pour déterminer les éléments pertinents pour la détection de la polarité , nous avons préalablement mené une analyse du corpus afin d'en extraire les caractéristiques les plus discriminantes pour chaque catégorie de tweet en polarité.

## 5 Analyse et mise en oeuvre

Nous allons ici présenter l’analyse exploratoire du corpus et ce qu’elle nous a permis d’ajouter comme règles ou comme traits pour l’apprentissage du modèle. Cette méthode sera par la suite mise à l’épreuve des tests sur corpus.

### 5.1 Analyse exploratoire

#### 5.1.1 Tweets émanant d’EDF

Une première interrogation sur le corpus a été l’impact des tweets émanant des comptes officiels liés à EDF, qui ont été conservés dans le corpus. Nous avons inclu tous les tweets venant de comptes liés à des entités EDF donc les comptes comme `edf_client`, et tous les comptes locaux comme `EDF_Tricastin`, `EDF_IDF`, `EDF_IsèreDrome` auxquels nous avons également ajouté les comptes associés aux personnalités liées à EDF comme par exemple le compte officiel de Jean Bernard Lévy. Ils représentent 9.5% du corpus positif et sont absents du corpus négatifs. Etant en dessous des 10%, on peut considérer qu’il n’y en a pas assez pour pouvoir réellement avoir des conséquences sur la construction du modèle.

#### 5.1.2 Vocabulaire spécifique

Les termes les plus spécifiques donnent généralement des indications sur les thèmes et les idées qui reviennent dans un corpus, nous avons cherché ces indications dans nos corpus à l’aide du logiciel de textométrie TXM<sup>5</sup>. Ils sont présentés en détail en annexe.

On retrouve dans le corpus positif, le vocabulaire de l’innovation (“innovation”, “smartgrid”) et le vocabulaire lié aux Prix Pulse, un “concours” annuel “fier” de faire “découvrir” et “accompagner” de nouvelles “startups” “innovantes”. On retrouve aussi un vocabulaire plutôt lié aux messages du service client (“Bonjour”, “merci”) qui est très spécifique au corpus positif même si les fréquences ne sont pas particulièrement importantes dans le corpus.

Les termes les plus fréquents du corpus négatif renvoient aux coupures de courant, nous observons des insultes et des thèmes d’actualité comme Linky qui a été très discuté dans les médias et sur les réseaux sociaux. Greenpeace étant des détracteurs récurrents d’EDF, ils sont aussi très cités. Enfin, nous observons des termes plutôt “positifs” comme “adorer” et “super”. Ces termes viennent des tweets ironiques concernant souvent les coupures de courant :

21. coupure de courant mdr super
22. Super la panne de courant générale
23. super y a une coupure de courant dans Hendaye
24. Plus d’électricité chez les dupdup super on adore
25. Coupure de courant on adore!

---

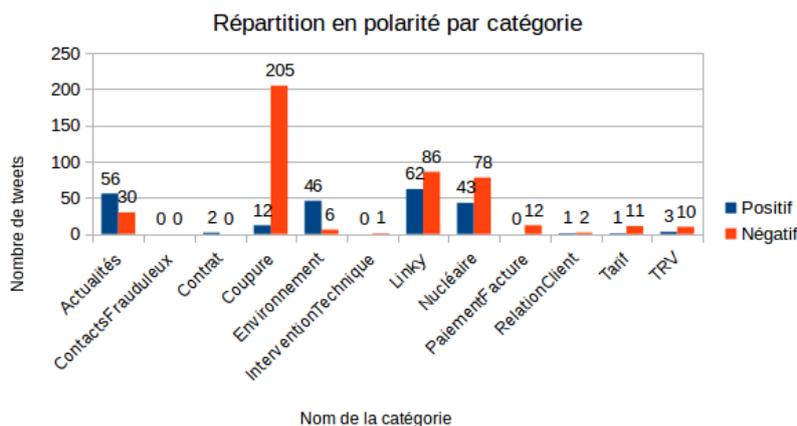
5. <https://halshs.archives-ouvertes.fr/halshs-00549779/fr/> et <https://halshs.archives-ouvertes.fr/halshs-00549764/en>

### 5.1.3 Particularités par catégorie

Une autre phase de l'exploration concerne les catégories et leur importance dans la détermination de la polarité d'un tweet. Ces catégories ont été attribuées en amont de cette étude, toute mauvaise classification n'est pas de notre fait. L'idée est de pouvoir utiliser cette information pour améliorer la classification en polarité.

Nous remarquons que les tweets non classés représentent respectivement 55% et 36% des tweets dans les corpus positif et négatif. Cependant, avec seulement 36% du corpus non-classé, une très claires majorités des tweets négatifs ont une catégorie. Nous cherchons ici à savoir si il est possible d'établir un lien fort entre catégorie thématique et polarité.

Il est à noter que certaines catégories ont une répartition de tweets positifs ou négatifs plus déséquilibrée que d'autres. Ces catégories ayant été mises au point pour permettre un repérage des tweets par thématique, il est intéressant ici d'étudier le lien entre thématique d'un tweet et polarité.



Prenons l'exemple de la catégorie "Environnement" qui contient 46 positifs contre 6 négatifs, on relève 3 tweets venant d'entités rattachées à EDF (ED-FEN1\_Officiel, erd1\_LR et 1.1\_Levy) et les tweets positifs mettent en avant principalement la politique de transition énergétique menée par EDF. Nous considérons donc cette catégorie comme discriminante.

La catégorie "Coupure" sera aussi un critère discriminant pour l'attribution d'une polarité, presque la totalité des tweets de cette catégorie sont négatifs.

Il est intéressant de constater que certaines catégories sont très majoritairement constituées de tweets en polarité, par exemple la catégorie "Coupure de Courant" est majoritairement négative. A l'inverse, les catégories "Actualités" ou "Linky" sont plus équilibrées.

Nous nous attarderons donc sur les catégories les plus "équilibrées", celle dans lesquelles le nombre de tweet positif ou négatif sont proches et ou la mise au point des règles pourra potentiellement aider à une meilleure classification.

**Actualités** Les tweets de la catégorie "Actualités" suivent différents schémas. On trouve d'abord les tweets positifs concentrés sur l'activité de la Fondation EDF et sur le Prix annuel Pulse, que l'on pourra considérer comme étant deux sous-thème de cette catégorie, ils se basent sur un lexique assez réduit :

26. EDF Pulse 2017 : inventons notre avenir électrique  
<https://t.co/Ps2hr07j7O> *positif*
27. #DigitalWeekEnd Idinvest et EDF s'associent pour investir dans les smart grids et la... <https://t.co/TKDWTNW8ZP>  
<https://t.co/1HA7CHP48T> *positif*
28. J'ai voté pour les Prix EDF Pulse! A votre tour de voter pour votre projet préféré! @EDFPulse <https://t.co/AXKz6K6KVO> *positif*
29. Comme moi, votez pour que Les Petites Gouttes remporte le Trophée du Public de 20 000€! <https://t.co/nmCZlsw2vL> #TDA2016 via @Fondation.EDF *positif*

Nous distinguerons ces tweets des "tweets d'information" (cf Guide d'annotation), relayant un fait qui pourra être présenté avec subjectivité ou non, avec un lexique plus large. Ceux-ci sont très présents dans cette catégorie. Nous constaterons rapidement qu'il n'y a pas de sous-thème identifiable pour ces tweets-là :

30. #EDF vend la moitié de #RTE à la Caisse des Dépôts  
<https://t.co/v5XyUsoZiz> via @Le\_Figaro *neutre*
31. EDF confirme sa décision d'investir dans deux réacteurs EPR à Hinkley Point <https://t.co/qahMmgWtdl>  
<https://t.co/J0TieRsUU3> *neutre*
32. Pour l'Anses, les compteurs Linky ne présentent pas de risque pour la santé - Le Monde <https://t.co/ml7ueYHBlj> *positif*
33. EDF plonge de 12% à la Bourse, l'exercice 2017 s'annonce difficile  
<https://t.co/n4wKCR1RBf> #AFP <https://t.co/NUBZl85rXO> *néгатif*

Il est à noter que cette catégorie se caractérise globalement par une présence très forte des déterminants.

Cette spécificité est prise en compte par l'analyse XIP puisqu'elle inclut une analyse morphosyntaxique de la phrase permettant de récupérer l'étiquetage de chacun des termes du tweet. Cependant, la surreprésentation de déterminants caractérise aussi bien les tweets positifs que les négatifs, ceux-ci étant très proches dans leur construction.

La mise au jour des sous-thématiques et leurs particularités reposant sur le lexique peuvent être utilisées comme éléments discriminants. XIP renvoie chacun des termes du tweets, la particularité des sous-thèmes est donc déjà prise en compte grâce à l'analyse de XIP. Cependant, l'étude des tweets de la catégorie Actualités ne permet pas d'ajouter des éléments discriminants supplémentaires pour la détection de la polarité des tweets.

**Linky** La mise en place des compteurs dits "intelligents" Linky a amené de nombreux débats, souvent houleux, sur Twitter. Entre relais d'informations par les journaux, organisation de réunions "anti-linky" dans certaines communes, mécontentement de certains usagers, promotion du nouveau compteur et de ces avantages puis désinformation sur la nocivité potentielle des "petits boîtiers verts" par d'autres, cette catégorie est relativement fournie en tweets divers et variés :

34. Linky : 1000 fois moins d'ondes qu'un frigo @enedis  
<https://t.co/Ix2aIv7TZa> <https://t.co/FP9zyE1Zf9> *positif*

35. Yaouuuuh!!! #SmartGrid en #P2P via #Blockchain! La France avance! merci @ecologiEnergie cc @InSunWeTrust @enedis <https://t.co/18R9IpfaId> *positif*
36. Déjà 20000 compteurs Linky dans l'Allier! Bravo à toutes les équipes. *positif*
37. Stop Linky à Grenoble : <https://t.co/Iy3D3F868Y> #replay *néгатif*
38. @CaptainTote Tout ça n'arrivera plus grâce au compteur Linky (mais tu auras le cancer et tu seras filmée par une caméra 360) *néгатif*
39. NON C FAUX BANDE DE MENTEURS ,, REFUSEZ TOUS LES LINKY ..... <https://t.co/sfToQNzWcf> *néгатif*

Dans cette catégorie, les tweets se divisent en deux grandes catégories, les "pro" et les "anti", respectivement associés à des tweets positifs et à des tweets négatifs.

Les premiers sont dans leur grande majorité, proches des tweets d'information. Ils se composent de relais de titres d'articles indiquant la non-dangerosité des compteurs pour les gens. Bien formés grammaticalement et comportant peu de fautes, ils renvoient aussi à des articles mettant en avant les économies d'énergie que permettent le nouveau compteur et la mise en avant des économies d'énergie qu'il permet.

Les "anti" se divisent eux en deux catégories. Il y a :

- d'une part les tweets se rapprochant aussi des tweets d'information, ils ont donc les mêmes caractéristiques que les tweets "pro" et se distinguent uniquement par leur vocabulaire. Les "anti" parlent d'"espionnage", des "dangers" des "ondes" et des abus des agents venus installés les compteurs contre l'avis des clients.
- d'autre part, des tweets plus proche de tweets détracteurs, des tweets rédigés en majuscules, contenant des insultes, une mauvaise construction et des erreurs de frappes et/ou d'orthographe.

Ces différentes caractéristiques ne sont cependant pas suffisantes pour rajouter des règles ou des éléments discriminants au modèle.

D'une part, nous avons des tweets "anti" proches des "tweets détracteurs" (tweets contenant des insulte), si bien qu'une partie de leurs particularités sont déjà prise en compte dans la caractérisation des tweet négatifs (prise en compte des majuscules, des insultes). Les erreurs, fautes de frappes, fautes d'orthographe, ne sont pas repérables ni par des règles, ni par des traits supplémentaires.

D'autre part, les "pro" sont proches de ce que nous avons appelé des tweets d'information. Ceux-ci, nous l'avons déjà expliqué, n'ont pas de caractéristiques suffisamment pertinentes pour pouvoir être repérés.

L'analyse de cette catégorie montre les différences entre le discours pour ou contre un événement ou ce que l'on pourrait qualifier de "nouvelle technologie" mais montre aussi l'appropriation du discours "pro" par les "anti".

Les tweets "pour Linky" et une partie des tweets "anti Linky" ont les mêmes caractéristiques, on ne peut donc pas se baser sur ces résultats pour former de nouvelles règles ou extraire des traits supplémentaires pour la détection de la polarité.

**Nucléaire** Le vocabulaire de la catégorie "Nucléaire" est composé de termes spécifiques aux centrales, "nucléaire", "réacteur" et "centrale", certains termes

comme “démantèlement”, “arrêt”, “ASN”, “indemnisation”, “cuve” et “anomalie”.

Les termes sont très souvent liés aux thèmes suivants qui sont trois points d’actualité : Flamanville, Hinley Point et Fessenheim.

Cette catégorie contient donc des tweets liés à l’actualité des centrales nucléaires mais encore une fois, elle ne permet pas une distinction nette dans la construction des tweets ou dans le vocabulaire pour faire une différence automatiquement entre les tweets positifs ou négatifs.

Beaucoup de militants anti-nucléaire seront plutôt présents dans la catégorie “Environnement” défendant le besoin de “maîtriser” la “consommation” énergétique ou de “réduire” l’utilisation du nucléaire et l’impact écologique de cette production d’énergie, enfin développer les “renouvelables”.

#### 5.1.4 Les tweets d’informations

Les tweets que nous avons appelé les “tweets d’information” ont montré leur omniprésence dans le corpus, à travers les catégories thématiques mais aussi de polarité. Ils ont mis en avant la difficulté à catégoriser et à annoter des tweets qui concernent des faits. Les tweets d’informations sont présents dans toutes les classes de polarité.

- Négatif
  - EDF va encaisser un milliard d’euros auprès de ses clients  
<https://t.co/abjyVfBTKS> via @Le\_Figaro
  - Face aux méthodes brutales pour imposer le compteur Linky, la résistance grandit... <https://t.co/q4lQamMLIO>  
<https://t.co/Jexqfxyv2H>
  - “EDF se suicide à Hinkley Point“ <https://t.co/audLPb7wLo>
- Positif
  - Comment EDF co-innove pour la transition énergétique  
<https://t.co/XMp0CMIUfe> via @upmagazine\_info
  - Nucléaire : feu vert d’EDF pour lancer la centrale d’Hinkley Point <https://t.co/Xk5no4p96N>
  - EDF veut aider les Français à produire leur électricité - L’offre d’autoconsommation, lancée jeudi 2 juin, marq...  
<https://t.co/u9A58NWaLV>
- Neutre
  - Compteurs Linky : quelles différences avec les compteurs électriques classiques ? <https://t.co/HZc7ZLuw2C>  
<https://t.co/jpReRvamfa>
  - Pourquoi EDF veut vendre la moitié de RTE  
<https://t.co/idbZ2dqK3F> via @Challenges
  - Fessenheim : EDF et l’Etat trouvent un compromis  
<https://t.co/oJ2KuB8tjR> <https://t.co/fRdBuoquX3>

De nombreux tweets se révèlent être des titres d’articles de journaux et la “syntaxe Twitter” se prête bien à un format court, abrégé, nécessitant d’aller à l’essentiel. Le corpus contient de nombreux tweets qui sont composés du titre de l’article et du lien vers l’article complet. Tous ces éléments font que l’on peut faire un parallèle entre les titres de journaux et les tweets.

Les éléments spécifiques aux titres de journaux ont été étudiés notamment dans “Syntaxe à la une : la structure des titres de journaux français et britan-

niques”1. Sont mis en avant notamment l’utilisation de termes argotiques et de structures syntaxiques abrégées et une surreprésentation des deux points. La syntaxe est caractérisée par l’omission de la copule ou de l’auxiliaire (40), une absence de verbe (41), ou une utilisation de l’infinitif (42), du présent (43) ou de l’impératif (44), (l’élément manquant est marqué par un A) :

- 40. Milosevic (A) seul face aux Alliés
- 41. Incendiaires (A) mis en examen
- 42. Des arbres à abattre dans le Gers
- 43. La SNCF lance les brigades ferroviaires
- 44. Voyagez en Yakoutie

Une des autres caractéristiques relevées est aussi l’utilisation de la forme passive avec omission de l’agent, l’utilisation de syntagmes nominaux sans article et de constructions du type Nom + Syntagme Prépositionnel :

- 45. Calvin Klein (A) jugé licencié
- 46. Mauvaises surprises pour les vacanciers,
- 47. Le manifeste des profs anti-Allègre.

L’interrogation est utilisée pour son effet rhétorique avec omission du syntagme verbal :

- 48. L’OM au stade de France ?

Enfin, une autre spécificité est la multiplication des voix, l’effacement du narrateur (l’auteur de l’article) et l’utilisation du discours rapporté :

- 49. Séguin : ”une élection n’est pas un pari”
- 50. Charles Pasqua affirme qu’il ”ne rentrera pas” au RPR.

Plusieurs de ces caractéristiques relèvent de l’absence d’un élément, la copule, l’auxiliaire ou le syntagme verbal, ce qui est difficile à repérer en utilisant des règles avec XIP.

### 5.1.5 Conclusion

D’une façon générale, nous avons pu constater que certaines catégories peuvent être liées à une polarité, et notons que certaines thématiques sont, en tant que tel, considérées comme positives ou négatives sur Twitter. L’utilisation de cette information, la thématique du tweet, est donc importante pour la détermination de la polarité. D’autres thématiques se sont révélées plus nuancées. Au sein de celles-ci, des sous-thèmes sont apparus et certains étaient de nouveau liables à une polarité. Nous observons donc de nouveau une corrélation entre un sujet et l’expression d’une opinion générale par rapport à ce sujet sur Twitter, l’adoption d’un avis général sur une thématique.

Du côté de la thématique de ”Linky”, nous avons vu la reprise par les ”anti-linky” de la rhétorique des ”pro-linky”, ce qui complique aussi une séparation nette dans la polarité et donc la mise en évidence de caractéristiques propres à chacun. Nous n’avons donc pas pu tirer de l’analyse de cette catégorie de caractéristiques supplémentaires pour la détection de la polarité.

Les tweets de type ”tweets d’information” couvrent beaucoup trop de thématiques, ils sont trans-catégories et trans-polarité. De plus, leur construction singulière,

comme par exemple l'absence de copule, est difficilement détectable avec des règles.

Au final, les informations tirées de l'analyse détaillées de ces catégories nuancées n'ont pas permis l'extraction de traits supplémentaires ou de motifs détectables par des règles.

## 5.2 Construction des règles

L'objectif des règles est de détecter des constructions ou des caractéristiques propres à un certain type de tweet afin de pouvoir créer des relations entre les mots ou groupes de mots. Ces relations sont ensuite utilisées comme features pour l'apprentissage automatique.

Par le biais de deux exemples, nous allons montrer la logique des règles que nous avons implémentées dans XIP.

De plus, dans les deux exemples ci-dessous, annotés come négatifs, l'objectif est de relever les éléments discriminants permettant de prédire la polarité du tweet.

51. Coupure de courant j'adore

52. Honte nationale, fardeau pour générations futures : "Nucléaire : EDF veut reporter le démantèlement"

Dans le premier, les relations renvoyées seront les suivantes : COUPUREC et IRONIE. La première relation COUPUREC détecte les constructions du type "coupure de courant". La règle utilisée pour renvoyer cette relation est la suivante :

$$\text{if}(\text{MOD-N}(\#1[\text{lemme :courant}], \#2)) \text{COUPUREC}(\#1, \#2).$$

Dans cette règle, nous cherchons une relation de type MOD-N (pour plus de détails sur une relation MOD-N, voir la description de la grammaire PASSAGE ici) ou modifieur du nom. Si nous détectons une relation MOD-N entre le lemme "courant" et un autre terme, nous créons une relation qui sera appelée COUPUREC entre les éléments tagués #1 ("courant") et #2 (le deuxième terme de la relation MOD-N qui a été repérée).

La deuxième relation, appelée IRONIE, se construit de la même façon en prenant en compte les traits de polarité définis dans le lexique :

$$\text{if}(\text{?}(\#1[\text{lemme :coupure}], \#2[\text{positif :+}])) \text{IRONIE}(\#1, \#2).$$

Nous cherchons ici n'importe quelle relation (ce qui est marqué par le ?) entre le terme "coupure" (#1[lemme :coupure]) et un mot ayant un trait "positif", sachant que le terme "adorer" possède le trait positif. Nous construisons ainsi une relation que nous avons nommée "IRONIE".

Pour le deuxième exemple :

Honte nationale, fardeau pour générations futures : "Nucléaire : EDF veut reporter le démantèlement"

Dans cet exemple, le but est de détecter la relation dans laquelle "honte" qui est un terme ayant un trait négatif, modifie "nationale". Le résultat étant une relation "honte nationale" qui sera considérée comme étant négative.

Cette règle est la suivante :

$$\text{if}(\text{MOD-N}(\#1, \#2[\text{négatif :+}])) \text{NEG\_MOD}(\#1, \#2).$$

XIP va renvoyer une relation de modifieur du nom entre "honte" et "nationale" de la forme : MOD-N(nationale,honte). Nous détectons cette relation et cherchons le trait "négatif" dans le deuxième terme de la relation, ici "honte". S'il est repéré, on crée une relation NEG\_MOD entre les deux termes. La même règle va aussi repérer le lien entre "fardeau" et "génération".

Nous pouvons détecter l'énumération de termes négatifs par le biais de la règle suivante :

```
if(JUXT(#1[negatif :+],#2[negatif :+])) NEG_REL(#1,#2).
```

Cette règle va renvoyer la juxtaposition entre "honte" et "fardeaux". Les énumérations se traduisent par des relations appelées JUXT dans XIP, nous nous attachons aux relations de juxtaposition entre deux termes négatifs pour créer une relation NEG\_REL.

### 5.3 Traits supplémentaires

Ces règles permettent d'extraire de XIP des traits qui vont être utilisés pour l'apprentissage automatique. Ils s'ajoutent à des traits de surface comme la détection du nombre de termes en majuscules, la catégorie du tweet, la ponctuation, et les émoticônes.

Nous incluons aussi un traitement préalable au passage des tweets dans XIP. Ce traitement comprend la suppression des adresses (@mon\_super\_compte\_twitter), la suppression des liens, le passage du tweet entièrement en minuscules et la suppression des #hashtags.

Ces pré-traitements sont minimes. Nous ne pouvons pas mettre au point une correction totale de chaque tweet, mais il était préférable d'enlever les liens car XIP fait une analyse par phrase, il est donc possible de lui donner plusieurs phrases à analyser, il va donc séparer en phrase sur la ponctuation (points d'exclamation, d'interrogation, point, etc). Les url sont composées de points (les liens twitter sont raccourcis en : http://t.co/...), l'analyse par XIP était donc faussée. De plus, les différents éléments de l'url sont considérés comme des mots et sont donc analysés tel quel par XIP, cela peut donc aussi fausser les relations trouvées au sein d'une phrase.

L'analyse de la phrase " Je crois que je sais où est le serveur de Facebook...??facebookdown https://t.co/5wB7UovJwJ" renvoie par exemple une relation :

```
MOD-N(https,facebookdown)
```

## 6 Résultats

### 6.1 Baseline

Afin de déterminer la méthode d'apprentissage automatique à utiliser, une première étape est d'établir une baseline. Ensuite nous comparerons les résultats obtenus avec différents algorithmes afin de voir lesquels seront les plus adaptés à notre tâche.

La baseline est à l'origine le résultat obtenu en effectuant une classification au hasard. Ici, nous nous appuyerons sur la définition communément admise dans les travaux de TAL, à savoir des résultats qui permettront une comparaison pertinente par rapport à l'évolution du système. Ce travail repose en partie sur l'outil XIP et la possibilité d'implémenter des règles. Il nous semble donc particulièrement intéressant (et pertinent) de prendre comme base de référence les résultats qui peuvent être obtenus sans XIP et de quantifier l'apport de l'analyse de XIP et des règles qui ont été écrites.

Nous prenons donc comme baseline les résultats obtenus avec un simple traitement des mots et de leur fréquence dans le corpus. Cela est implémenté dans la librairie scikit-learn sous la forme d'une méthode : `CountVectorizer`<sup>6</sup>. Cette méthode effectue en prétraitement une tokenization et un filtrage des mots vides. Il construit ensuite un vecteur par document (ici par tweet).

Nous avons testé cette baseline sur plusieurs algorithmes afin de déterminer lequel serait le plus adapté à notre tâche. Les algorithmes testés et leurs résultats sont donnés dans le tableau suivant. Celui-ci résume seulement les mesures d'exactitude obtenus, les résultats complets sont disponibles en annexe.

Algorithmes	Positif	Négatif
Régression logistique	0.718	0.721
Bayes multinomial	0.718	0.658
Bayes gaussien	0.638	0.647
K plus proches voisins	0.624	0.571
Arbres de décision	0.591	0.665
SVM	0.707	0.718

La régression logistique donne les meilleurs résultats et est suivie par les SVM. Nous avons au final retenu les SMV qui offrent de meilleurs résultats lors de l'ajout de features supplémentaires.

### 6.2 Apport de XIP

Pour avoir une première idée de l'apport de XIP dans la classification, nous avons testé, avec des SVM, les différences de résultat incluant ou excluant les résultats de XIP. Dans le tableau suivant, les colonnes "sans XIP" résultent d'un apprentissage uniquement sur les mots et leur fréquence dans l'ensemble du corpus (baseline). Les chiffres des colonnes "XIP" correspondent aux résultats d'un apprentissage avec comme features les éléments constituant la sortie de XIP. Enfin, les colonnes "Règles" sont les résultats obtenus en utilisant XIP et la grammaire mise au point avec le lexique et les règles.

6. Détails disponibles sur le site de sk-learn : <http://scikit-learn.org>

<b>Baseline</b>			
	Précision	Rappel	F-mesure
Neutre	0.70	0.72	0.71
Négatif	0.73	0.71	0.71
Total	0.71	0.71	0.71
<b>XIP</b>			
	Précision	Rappel	F-mesure
Neutre	0.64	0.72	0.68
Négatif	0.70	0.62	0.66
Total	0.67	0.67	0.67
<b>XIP + Règles</b>			
	Précision	Rappel	F-mesure
Neutre	0.76	0.66	0.71
Négatif	0.72	0.81	0.76
Total	0.74	0.74	0.74

Pour les tweets négatifs, on arrive à obtenir un rappel relativement satisfaisant mais la précision est insuffisante.

L'apport des règles sur la détection est non négligeable alors que les résultats de XIP seuls intégrés au modèle font baisser les résultats pour la détection des tweets négatifs. Les règles créent différentes relations, celles-ci font partie des features ayant les scores les plus importants pour le modèle, preuve de leur efficacité. Cependant, les résultats montrent qu'un affinement des règles pourrait permettre d'aller plus loin et d'améliorer les performances, notamment en termes de précision.

<b>Baseline</b>			
	Précision	Rappel	F-mesure
Neutre	0.73	0.70	0.71
Positif	0.67	0.71	0.69
Total	0.70	0.70	0.70
<b>XIP</b>			
	Précision	Rappel	F-mesure
Neutre	0.80	0.51	0.62
Positif	0.61	0.86	0.71
Total	0.71	0.67	0.67
<b>XIP + Règles</b>			
	Précision	Rappel	F-mesure
Neutre	0.82	0.88	0.84
Positif	0.85	0.78	0.81
Total	0.83	0.83	0.83

Sur les tweets positifs, XIP est performant et permet une nette amélioration des résultats. Les règles améliorent globalement les performances et la détection des tweets neutres, mais on voit une baisse du rappel dans la détection des tweets positifs.

Cela s'explique par une moindre prise en compte du lexique et un score élevé pour les relations créées par les règles. En ajoutant dans la grammaire positive, les relations pour la détection des tweets traitant de Linky, on constate une baisse du rappel mais une augmentation de la précision. De même, certaines

règles concernant la détection des tweets négatifs sont ajoutées à la grammaire "positive" et cela participe à améliorer les résultats.

En conclusion, les tweets négatifs sont trop largement détectés, si bien qu'une partie des tweets annotés négatifs sont en fait neutres, tandis que les tweets positifs sont moins souvent détectés mais de façon sûre. En conclusion, les règles sont trop précises pour les tweets positifs et pas assez pour les négatifs.

## 6.3 Réseaux de neurones

Nous avons aussi testé un modèle avec un Réseau de Neurones Convolutif et Word2vec pour la vectorisation des tweets. Les résultats sont présentés ci-après.

### 6.3.1 Modèle négatif/positif

Nous avons avant tout testé une classification en négatif/positif. Comme le système des réseaux de Neurones ne nous renvoie pas une classe mais un score (une probabilité) d'appartenance à une classe, il était intéressant de tester et d'étudier la possibilité de faire une classification en fonction de ce score.

	Précision	Rappel	F-mesure	Support
Négatif	0.84	0.72	0.78	65
Positif	0.68	0.81	0.74	47
Total	0.77	0.76	0.76	112

Il s'est avéré que les scores obtenus sont soit très élevés (0,99) soit très très bas ( $10 \times 10^{-5}$ ). Sachant que le nombre de tweets neutres (prépondérant dans le corpus) qui devraient obtenir un score moyen, cette méthode de semble pas être suffisamment satisfaisante.

Nous avons tout de même essayé de construire deux modèles séparés, un pour la détection des tweets négatifs et un pour la détection des tweets positifs.

### 6.3.2 Modèle négatif/neutre

	Précision	Rappel	F-mesure	Support
Négatif	0.61	0.79	0.69	76
Neutre	0.88	0.75	0.81	150
Total	0.79	0.76	0.77	226

### 6.3.3 Modèle positif/neutre

	Précision	Rappel	F-mesure	Support
Négatif	0.58	0.79	0.67	52
Neutre	0.93	0.83	0.87	173
Total	0.85	0.82	0.83	225

Ces résultats sont relativement satisfaisants mais nous remarquons qu'ils sont dus à une bonne détection des tweets neutres. Or nous nous intéressons en priorité aux tweets en polarité. Les mesures sur les tweets en polarité sont légèrement inférieures à ce que l'on obtient avec des algorithmes de machine learning et XIP.

## 7 Conclusion

### 7.1 Croisement des résultats pour l'intégration dans Tweets-Tracker

L'objectif de ce travail était la mise en place d'une détection plus fine de l'opinion sur Twitter en vue d'une intégration dans l'application existante Tweets-Tracker.

Afin de réaliser cette intégration, il était nécessaire de croiser les résultats. La création de deux modèles séparés pour une amélioration de la détection de la polarité demande un recoupement de ces résultats.

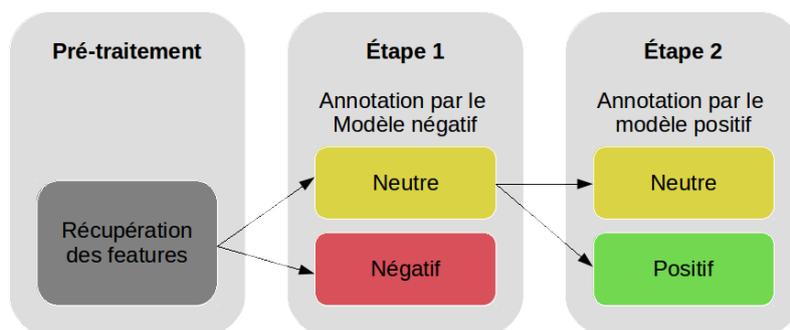
Nous avons d'abord pensé faire une validation double. Cela signifie qu'un tweet sera annoté par les deux modèles en parallèle et qu'on comparera les deux annotations. Plusieurs cas sont alors possibles. Les colonnes du tableau suivant montrent les annotations du modèle positif, un tweet est annoté soit positif, soit neutre. Les lignes montrent les annotations possibles par le modèle négatif, les tweets sont soit négatifs soit neutres. Nous pouvons donc observer les différentes possibilités de croisement des annotations.

	Neutre	Positif
Neutre	Neutre (1)	Positif(2)
Négatif	Négatif (2)	? (3)

Le premier cas (1) est celui où les deux modèles prédisent une annotation neutre, dans ce cas le tweet conserve une annotation neutre, les deux modèles sont d'accord.

Dans le cas d'une annotation neutre par l'un des modèles et d'une annotation en polarité par le deuxième modèle, c'est l'annotation en polarité qui est conservée, cas (2). Enfin, le problème concerne principalement le cas (3) où les deux modèles font une annotation chacun dans leur polarité.

On peut aussi choisir de donner la priorité à un modèle. Un nouveau corpus de test a été constitué afin de déterminer la meilleure combinaison d'enchaînement des deux modèles.



Après avoir passé les pré-traitements nécessaires aux deux modèles (étape 1), le tweet sera d'abord annoté par le modèle négatif (étape 3). Si le tweet est annoté négativement, il garde cette étiquette, sinon il sera annoté par le modèle positif. A cette étape (étape 3), si le tweet annoté neutre par le modèle négatif, est annoté neutre par le modèle positif, il restera donc neutre. A l'inverse, s'il est annoté positif par le modèle positif, il deviendra positif.

Les différents tests ont montré que la plupart des erreurs viennent des confusions entre les tweets négatifs et neutres. Le nombre de tweets annotés à la fois positif et négatif par les deux modèles ne sont pas majoritaires et plutôt à la marge. La principale différence entre les deux systèmes, chacun mettant ne priorité l'un ou l'autre modèle, montrent que ce sont ces tweets, principalement neutres, qui font la différence.

## 7.2 Des thématiques intrinsèquement liées à une polarité

Nous avons pu constater une corrélation entre certaines catégories (chacune correspondant à une thématique) et une polarité. Les tweets traitant des coupures de courant en sont l'exemple le plus marquant, au sein de notre corpus, sur 433 tweets, 47% sont négatifs et 0.4% sont positifs.

L'analyse des catégories a aussi mis en évidence la présence de sous-thèmes qui eux aussi pouvaient être liés à une polarité, par exemple la Fondation EDF dans la catégorie Actualités. On peut donc associer certaines thématiques à une polarité et voir dans ce lien l'expression d'une opinion générale favorable ou non à un sujet.

Nous avons aussi vu la forte présence de ce que nous avons appelé les "tweets d'information" qui est un type de tweet qui se démarque par ses sujets multiples et son expression dans toutes les polarités. L'utilisation répétée de ce tweet en fait un élément important dans la détection de la polarité.

Au vu de leur construction que nous avons détaillé ici, la possibilité d'en faire une analyse plus poussée apparaît intéressante. Nous sommes ici face à un type de tweets qui semble demander un traitement particulier de part leur prépondérance dans le corpus et leur présence trans-catégorie et trans-polarité. Ces tweets se sont révélés être les tweets les plus complexes à prendre en compte et nous avons exposé les difficultés à mettre au point des règles qui permettent de les classer en polarité et de les distinguer automatiquement.

Cette étude pose donc la question d'un changement d'angle dans la recherche de la polarité sur Twitter et interroge la possibilité de partir d'une recherche de thématiques, lesquelles seront ensuite liées à une polarité en fonction de l'expression des réseaux sociaux à leur sujet.

Le suivi de l'évolution de cette opinion générale sera possible via les moyens actuels qui peuvent donner des résultats satisfaisants sur un corpus restreint en termes de typologie de tweets et de thèmes. Twitter est aujourd'hui spécifiquement complexe à étudier du fait de la multiplicité des types de tweets et des thèmes abordés. Une telle méthode pourrait permettre d'éviter ces deux écueils.

## 7.3 Evolution des thématiques sur Twitter

La particularité de Twitter repose aussi sur sa réactivité aux sujets d'actualités et à ses changements, à ses évolutions en fonction des idées et des modes ou des innovations. Il est difficile de ne pas prendre en compte les possibilités de changement qui sont intrinsèques à ce réseau social et à son fonctionnement.

Une évolution de ces thématiques est attendue mais certaines sont susceptibles de rester stables. Les thématiques récurrentes qui s'inscrivent dans le long terme notamment dans le contexte EDF sont nombreuses. Même si nous pouvons prédire une évolution de certaines thématiques, comme la question des compteurs Linky qui sont en débat depuis leur création, d'autres sont amenés à

durer, à s'inscrire dans le temps. La catégorie "Coupure" ne connaîtra à priori pas de changements quant à sa structure. Même si la catégorie peut perdre en importance (proportionnellement à la baisse des fréquences des coupures de courant par exemple), elle continuera à être alimentée et nous pouvons attendre une continuité dans le vocabulaire.

Il est cependant possible que l'opinion générale sur une catégorie, que nous avons qualifiée de "mitigée" aujourd'hui, soit amenée à évoluer. Il est possible que d'ici 5 ou 10 ans, les compteurs Linky prouvent leur efficacité, leur praticité et soient loués par tous les utilisateurs pour les économies d'énergie apportées par exemple, ou l'inverse.

Ainsi, les évolutions peuvent impacter les thématiques, dans le cas par exemple de l'apparition d'une nouvelle thématique, ou la polarité par le biais d'un changement du sentiment général face à une thématique.

## Références

- [Audrey and Boullier, 2012] Audrey, L. and Boullier, D. (2012). *Opinion Mining et Sentiment analysis, Méthodes et outils*.
- [Bo and Lillian, 2008] Bo, P. and Lillian, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2) :1–135.
- [Brun, 2011] Brun, C. (2011). Un système de détection d’opinions fondé sur l’analyse syntaxique profonde.
- [Brun et al., 2014] Brun, C., Nicoleta Popa, D., and Roux, C. (2014). Xrce : Hybrid classification for aspect-based sentiment analysis.
- [Brun and Roux, 2014] Brun, C. and Roux, C. (2014). Decomposing hashtags to improve tweet polarity classification (décomposition des ” hash tags ” pour l’amélioration de la classification en polarité des ” tweets ”) [in french]. pages 473–478.
- [Cori, 2008] Cori, M. (2008). Des méthodes de traitement automatique aux linguistiques fondées sur les corpus. 171.
- [Go et al., 2009] Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. 150.
- [Hamon et al., 2015] Hamon, T., Fraise, A., Paroubek, P., Zweigenbaum, P., and Grouin, C. (2015). Analyse des émotions, sentiments et opinions exprimés dans les tweets : présentation et résultats de l’édition 2015 du défi fouille de texte (DEFT). In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2015)*, Caen, France.
- [Maurel et al., 2017] Maurel, S., Curtoni, P., and Dini, L. (2017). Classification d’opinions par méthodes symbolique, statistique et hybride.
- [Pak and Paroubek, 2011] Pak, A. and Paroubek, P. (2011). Classification en polarité de sentiments avec une représentation textuelle à base de sous-graphes d’arbres de dépendances. In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles*, Montpellier, France. Association pour le Traitement Automatique des Langues.

## 8 Annexes

### A Scripts d'extraction des tweets

Script d'extraction des tweets à partir de la base de données elasticsearch

```
#!/ coding: utf-8

'''
extraction de 2250 tweets
entr e : requete elasticsearch
sortie :
    corpus_tweets.txt (tabulations) :
        id_tweet, date, nom_utilisateur, categorie, tweet
    corpus_tweets.xlsx
        id_tweet, date, nom_utilisateur, categorie, tweet
    hashtags.txt : liste des hashtags et nombre
        d'appararition (ordre selon le nombre d'apparition)
    emojis.txt : liste des emojis et nombre
        d'apparition (ordre selon le nombre d'apparition)
'''

from elasticsearch import Elasticsearch, helpers
import json
import sys
import random
from pprint import pprint
from openpyxl import Workbook
from openpyxl.styles import Border, Alignment, Font,
    Side

if __name__ == '__main__':
    inpath = "../input/"
    outpath = "../input/"

    # initialisation des variables
    ref_cats = {
        "Actualites" : 169,
        "ContactsFrauduleux" : 2,
        "Contrat" : 8,
        "Coupure" : 433,
        "Environnement" : 84,
        "InterventionTechnique" : 8,
        "Linky" : 257,
        "Nucleaire" : 301,
        "PaiementFacture" : 21,
        "RelationClient" : 8,
        "Tarif" : 24,
```

```

        "TRV" : 21,
        "arcachon" : 1
    }
    cats = {
        "Actualites" : 0,
        "ContactsFrauduleux" : 0,
        "Contrat" : 0,
        "Coupure" : 0,
        "Environnement" : 0,
        "InterventionTechnique" : 0,
        "Linky" : 0,
        "Nucleaire" : 0,
        "PaielementFacture" : 0,
        "RelationClient" : 0,
        "Tarif" : 0,
        "TRV" : 0,
        "arcachon" : 0,
        'noclass' : 0
    }

    dico_hash = {}
    dico_emojis = {}
    tweets_list = {}
    liste_id = []
    cpt_tweets = 0
    esfrom = 0

# connexion Elasticsearch
    es = Elasticsearch([
        {'host': 'hcc08node.acc.edf.fr', 'port': 9200},
    ])

    body = {
        "size":1000,
        "sort": {
            "fieldEDF_date":{ "order":"asc" }
        },
        "query":{ "match":{ "fieldEDF_isRT": False }},
        "filter" : {
            "bool" : {
                "must" : {
                    "range" : {
                        "fieldEDF_date": {
                            "gte": "2016-05-01T00:00:00",
                            "lte": "2017-05-01T00:00:00"
                        }
                    }
                }
            }
        }
    }

```

```

    }
}

# dans les resultats de la recherche :
# recuperer les tweets, leur id,
# l'auteur, les categories ...
# creer une liste des tweets recuperes
# transformer en fonction qui cree un generateur
while True:
    body["from"] = esfrom
    res = es.search(index = "tweetstream",body=body)
    hits = res["hits"]["hits"]
    if len(hits) == 0 :
        break

    for tweet in hits:
        cat = ""
        id_tweet = tweet['_source']['fieldEDF_id']
        date = tweet['_source']['fieldEDF_date']
        auteur = tweet['_source']['fieldEDF_auteur']
        message = tweet['_source']['fieldEDF_tweet'].encode('utf-8')
        tweets_list[id_tweet] = {}
        tweets_list[id_tweet]['id_tweet'] = id_tweet
        tweets_list[id_tweet]['date'] = date
        tweets_list[id_tweet]['auteur'] = auteur
        tweets_list[id_tweet]['tweet'] = message
        if 'fieldEDF_cat' in tweet['_source']:
            cat = tweet['_source']['fieldEDF_cat']
        tweets_list[id_tweet]['cat'] = cat
        liste_id.append(id_tweet)

    print(esfrom)
    esfrom += 1000
    # on recupere les tweets par paquet de 1000

# => on a recupere l'ensemble des tweets et
# leurs informations dans un dictionnaire tweets_list
# de la forme :
# {id:{'id':id_tweet,'date':date,
# 'cat':string_des_categories,'tweet':contenu_du_tweet},{id: ...}}

# impression du nombre de tweets
print(len(tweets_list))

# traitement et impression des tweets
output_name = "corpus_positif_der_der"
output = open(inpath+output_name+".txt","w")

```

```

wb = Workbook()
ws = wb.active

cpt_boucle = 0
for element in tweets_list:
    # max tweet atteint ?
    egalite = True
    ref_cats['noclass'] = cats['noclass']
    for categorie in cats:
        if cats[categorie] == "noclass":
            continue
        if cats[categorie] != ref_cats[categorie]:
            egalite = False
            break
    # si toutes les cat gories sont remplies et
    # si on a 2250 tweets alors fin de la boucle
    if egalite == True and cpt_tweets >= 2250:
        break

# cat = noclass ?
if tweets_list[element]["cat"] == "":
    if cats['noclass'] > 1002:
        # passer au tweet suivant
        continue
    else:
        # AJOUT TWEET
        cats["noclass"] += 1
        tweet = tweets_list[element]['tweet']
        output.write(unicode(tweets_list[element]['id_tweet']))
        output.write(u"\t")
        output.write(unicode(tweets_list[element]['date']))
        output.write(u"\t")
        output.write(unicode(tweets_list[element]['auteur']))
        output.write(u"\t")
        output.write(unicode(tweets_list[element]['cat']))
        output.write(u"\t")
        output.write(tweet)
        output.write(u"\n")

        infos = []
        infos.append(tweets_list[element]['id_tweet'])
        infos.append(tweets_list[element]['auteur'])
        infos.append(tweets_list[element]['date'])
        infos.append(tweets_list[element]['cat'])
        infos.append(tweets_list[element]['tweet'])
        ws.append(infos)
        cpt_tweets+=1
        continue

# cat = une cat

```

```

for cat in tweets_list[element]['cat'].split("_"):
    # cat in cats ?
    # max cat atteint ?
    if cats[cat] >= ref_cats[cat]:
        # on sort de la boucle
        break
    # si max cat pas atteint
    else:
        for cat in tweets_list[element]['cat'].split("_"):
            if cats[cat] == 'noclass':
                continue
            if cats[cat] < ref_cats[cat]:
                cats[cat] += 1
            else:
                break
        tweet = tweets_list[element]['tweet']
        output.write(unicode(tweets_list[element]['id_tweet']))
        output.write(u"\t")
        output.write(unicode(tweets_list[element]['date']))
        output.write(u"\t")
        output.write(unicode(tweets_list[element]['auteur']))
        output.write(u"\t")
        output.write(unicode(tweets_list[element]['cat']))
        output.write(u"\t")
        output.write(tweet)
        output.write(u"\n")

        infos = []
        infos.append(tweets_list[element]['id_tweet'])
        infos.append(tweets_list[element]['auteur'])
        infos.append(tweets_list[element]['date'])
        infos.append(tweets_list[element]['cat'])
        infos.append(tweets_list[element]['tweet'])
        ws.append(infos)
        cpt_tweets+=1
        break # NEXT TWEET

    # pprint(cats)

# criture et sauvegarde du fichier excel
print("Cr ation du fichier "+outpath+output_name+".txt")
print("Cr ation du fichier "+outpath+output_name+".xlsx")
wb.save(outpath+output_name+".xlsx")

pprint(cats)
output.close()

```

## **B Guide d'annotation**

# PROJET DE TRAVAIL SUR LA POLARITE DE MESSAGES ISSUS DE TWITTER

---



## Annexe 1 : Guide d'annotation

Mai 2017

Rédacteur : Mathilde POULAIN



# SOMMAIRE

---

<b>PRESENTATION</b>	<b>3</b>
<b>LES DONNEES</b>	<b>4</b>
1. CONSTITUTION DU CORPUS	4
2. REPARTITION DE L'ANNOTATION	5
<b>CONSIGNES D'ANNOTATION</b>	<b>6</b>
1. QU'EST-CE QU'UN TWEET POSITIF ?	6
2. QU'EST-CE QU'UN TWEET NEGATIF ?	6
3. QU'EST-CE QU'UN TWEET NEUTRE ?	7
<b>PRECISIONS, CAS PARTICULIERS</b>	<b>9</b>
1. LES TWEETS DITS « D'INFORMATION »	9
2. LES TWEETS SARCASTIQUES OU IRONIQUES	10
<b>EXEMPLES D'ANNOTATION</b>	<b>11</b>

# PRESENTATION

---

Le volume des données numériques textuelles, disponibles sur Internet (forums, twitter etc.) augmente chaque année. L'analyse de ces informations, structurées ou non, est, aujourd'hui, un impératif stratégique pour une entreprise telle qu'EDF.

A l'ère du tout numérique, il devient stratégique d'exploiter l'expression spontanée issue des réseaux sociaux dans des délais de plus en plus courts. Ces derniers sont d'autant plus importants qu'aujourd'hui ils représentent une mine d'informations précieuses.

Dans cette optique, le projet consiste à analyser et restituer l'expression des réseaux sociaux sur EDF en mettant en place un système d'annotation automatique en polarité.

Ce guide d'annotation présente dans un premier temps les données qui seront traitées, le corpus qu'elles composent et la répartition des tweets par annotateurs. Ensuite, les consignes d'annotation sont détaillées à travers la définition et l'explicitation des différentes catégories, et enfin quelques précisions sont données sur des cas particuliers ou pouvant porter à confusion.

# LES DONNEES

---

## 1. Constitution du corpus

L'application TweetsTracker a permis l'extraction d'un corpus conséquent qui a été dans un premier temps classifié en fonction du sujet global du tweet. Les catégories sont les suivantes :

- Actualités
- Démarchages frauduleux
- Contrat
- Coupure
- Intervention technique
- Paiement facture
- Relation client
- Tarif
- TRV
- Linky
- Nucléaire

Certains messages, ne rentrant dans aucune catégorie, sont non classés et n'ont pas d'étiquette.

Les tweets sont récupérés avec les informations suivantes :

- un identifiant
- la date
- le nom de l'auteur (ou pseudo),
- les URLs,
- le lemme de chaque terme,
- le tweet lemmatisé,
- l'information si c'est un retweet,
- la catégorie attribuée
- le texte du message.

Dans le corpus constitué pour ce projet, la répartition des tweets par catégorie est respectée, on reproduit donc dans cet échantillon la même distribution par catégorie que dans l'ensemble des tweets stockés afin d'avoir un ensemble aussi homogène dans sa globalité.

Pour chaque tweet, un prétraitement aura été effectué. Les messages accompagnés d'URLs sont très nombreux mais ces liens ne seront pas pris en compte lors de l'analyse, ils auront donc été retirés au préalable. Les hashtags et noms d'utilisateur seront, par contre, laissés dans l'état.

## 2. Répartition de l'annotation

Pour un corpus total de 2250 tweets, 7 annotateurs se verront attribuer 500 tweets chacun plus une personne qui annotera 1000 tweets.

Afin d'avoir une double annotation, chaque ensemble de 500 tweets sera divisé en deux parties de 250 messages qui seront chacune annotée par deux annotateurs.

Ainsi, pour chaque annotateur, sur l'ensemble de ses 500 tweets à annoter, la première partie sera annotée par une deuxième personne, et la deuxième partie sera annotée par une troisième personne.

Annotateurs	A1	A2	A3	A4	A5	A6	A7	A8
E1	✓	✓						
E2	✓		✓					
E3	✓			✓				
E4	✓				✓			
E5		✓				✓		
E6			✓	✓				
E7					✓		✓	
E8						✓		✓
E9							✓	✓

L'extrait 1 sera annoté par les annotateurs A1 et A5, l'extrait 2 sera annoté par les annotateurs A1 et A6 ...

Cela correspond au final à 9 ensembles de 250 tweets, soit un corpus final de 2250 tweets avec une double annotation.

# CONSIGNES D'ANNOTATION

---

Un tweet peut être annoté comme positif, négatif ou neutre. Chaque tweet est considéré comme un ensemble, il n'y a pas de découpage interne du message, il faut considérer la polarité générale du tweet.

Le tweet est annoté au regard du sujet dont il parle, ce sujet peut être EDF en général, ou tout élément relatif à EDF, son service client, ses tarifs ou par exemple les coupures de courant qui lui sont imputées.

L'annotation double n'est pas possible, chaque message ne peut appartenir qu'à une et seulement une catégorie. En cas de doute ou de tweet pour lequel deux opinions sont clairement exprimées (sans être ironique), alors le tweet rentrera dans la catégorie neutre. Il n'y a pas de catégorie « non classés ».

## 1. Qu'est-ce qu'un tweet positif ?

Un tweet est annoté comme positif s'il expose des sentiments positifs (joie, amusement, réjouissance, reconnaissance ...), une information positive, ou s'il présente une information d'un point de vue « approbateur » ou favorable :

- Découvrez comment #EDF utilise le #Numérique au service de la mobilité interne ! #TransfoNum <https://t.co/jnBrgtitaa> <https://t.co/TEt4sFWAQJ>
- Création de la société #Edvance : la refondation du #nucléaire français avance. #edf #areva <https://t.co/leSsViAaY7>
- Bravo @edf\_client pour la qualité du contact obtenu lors de mon appel ! T pro, ça change #pourvuquecadure #edf #serviceclients
- La conseillère EDF elle est trop gentille .. Coeur sur elle 🍷 @edf\_client #Mary

On inscrit dans les tweets positifs les messages traitant de sujets comme l'innovation, la mise en place de solutions modernes, le renouvellement, ou encore les énergies renouvelables et l'écologie qui sont des sujets dominant dans la twittosphère EDF.

## 2. Qu'est-ce qu'un tweet négatif ?

On considère un tweet comme négatif lorsqu'il présente une opinion négative ou émet un commentaire négatif sur un sujet en lien avec EDF. On considère aussi dans cette catégorie les faits ou informations présentées selon un point de vue réprobateur.

- y'a une coupure de courant chez moi putain ☐
- Les 7 raisons de refuser les compteurs communicants Gazpar, Linky et cie...  
<https://t.co/1HTgXY5vdr>
- La facture EDF qui fâche
- Le ventilateur qui marche H24, on va rigoler quand on va recevoir la facture d'EDF.

On inclut donc dans les tweets négatifs les messages présentant une information (des faits dits objectifs) selon un point de vue ou une opinion défavorable. Dans l'exemple suivant, l'information est objective : « 100% d'électricité renouvelable pas plus cher que le nucléaire » mais elle est présentée de façon négative comme étant une « étude qui dérange » :

- « Une étude qui dérange : 100% d'électricité renouvelable pas plus cher que le nucléaire ! | @scoopit <https://t.co/L8KwRBDxmA> »

Le tweet sera annoté comme négatif, nous détaillons le cas particulier de l'annotation des « tweets d'information » [plus loin](#).

En règle générale, un tweet concernant les conséquences négatives d'une action relevant d'EDF ou d'une de ses entités sera aussi considéré comme négatif, cela concerne par exemple les tweets à propos de coupures de courant :

- Montparnasse: circulation des TGV perturbée par une panne électrique  
<https://t.co/MAxFBBOSBF> <https://t.co/gxxC0KEd0I>

Dans cet exemple, le tweet traite de la conséquence négative d'une coupure de courant, bien que la référence à une coupure de courant ne soit pas automatiquement négative, elle est ici liée aux « dommages » causés par celle-ci, ce tweet sera donc annoté négativement.

### 3. Qu'est-ce qu'un tweet neutre ?

Un tweet neutre présente généralement des faits et les expose sans prendre un point de vue particulier ni présenter une argumentation pour ou contre les faits exposés.

- @edf\_client bonjour j'ai reçu une relance de 35.04€ j'ai posté le chèque le 16/05 Que faire ? Merci
- EXCLUSIF. Une trentaine de consultants d'EDF perquisitionnés <https://t.co/bSd5jiikWv> via @LePoint
- Juste pas de coupure d'électricité stp l'orage
- @EU\_Commission autorise le rachat de l'activité réacteurs nucléaires du Areva par EDF  
<https://t.co/anE>
- Quelqu'un peut me dire si le compteur linky est obligatoire ou non ? ☹

Par défaut, on considère que les tweets traitent d'EDF, si toutefois certains tweets ne semblent pas avoir pour objet EDF ou un élément qui lui est lié, ces messages sont aussi annotés comme « neutre ». Par exemple :

- Coupure De Courant by Françoise Hardy <https://t.co/Qk4niw8tbl> #CoupureDeCourant #Fr
- Probablement pas de stream ce soir parce que SFR refuse de venir réparer la panne si je n'appelle pas, alors qu'ils sont au courant du pb

Ces tweets ne relèvent pas d'EDF et appartiennent donc à la catégorie neutre, peu importe leur polarité.

De même, les tweets hors sujet, comme les tweets humoristiques ou moqueurs que nous détaillons [plus loin](#), sont à inclure dans la catégorie neutre.

# PRECISIONS, CAS PARTICULIERS

---

## 1. Les tweets dits « d'information »

Nous nous attardons dans cette partie sur ce que nous avons appelé les « tweets d'information », on englobe dans ce cadre tout message qui aura généralement été posté par des organes d'informations comme les journaux ou magazines d'informations et qui présente une actualité. Cela concerne également les tweets émanant de particuliers s'ils font circuler des liens vers des articles d'actualité ou les messages issus de sites de ré-information. Nous détaillons ici l'annotation pour ces tweets particuliers, souvent à mi-chemin entre de l'information objective et une prise de position par rapport à cette information.

Une information négative ou positive présentée de façon neutre reste une information négative ou positive, elle est donc annotée comme tel :

- « Manifestations contre Linky, Gazpar ont lieu dans toute la France » est annoté négativement.

Les informations à priori objectives présentées selon un point de vue et avec une prise de position pour ou contre est annotée selon la prise de position de l'auteur :

- « Comprenez-vous que l'on débaptise tous nos fleurons nationaux (TGV, EDF etc.) pour y accoler des noms anglais ? » est négatif

Une information peut donc **être** objective ou orientée et **être présentée de façon** objective ou orientée. Pour résumer et illustrer cette problématique, prendre en considération l'exemple suivant :

- « Une panne électrique dans un datacenter de British Airways provoque le chaos » est annoté comme négatif

versus

- « British Airways : panne de courant ou externalisation mal maîtrisée ? » qui sera annoté neutre.

Dans le premier cas, l'information est présentée de façon négative, par conséquent, le tweet est annoté comme négatif. Dans le second cas, l'information est objective (elle ne critique pas ni ne remet en cause EDF) et présentée de façon neutre, le tweet est donc annoté comme « neutre ». Une même information peut donc être annotée de deux façons différentes selon la polarité du message qui l'accompagne.

## 2. Les tweets sarcastiques ou ironiques

Nous considérons dans cette partie, les tweets railleurs ou moqueurs comme :

- « J'entends les rageux qui ont plus d'électricité, mdrrrr »
- « Tiphaine en pls en message à cause de sa panne de courant, un délire »

Ces tweets ne contiennent pas de plaintes contre EDF et ne sont donc pas à annoter comme « négatifs », ils ne concernent pas un avis ou une opinion sur EDF ou ses services.

Les tweets au ton ironique **concernant EDF** seront par contre étiquetés selon l'avis qu'ils présentent :

- « Coupure de courant, j'adore j'adhère je valide » sera annoté négativement (ironie)

# EXEMPLES D'ANNOTATION

Rappel des exemples explicités dans les parties précédentes et exemples supplémentaires sur l'annotation de la polarité :

Texte du tweet	Polarité
TILLAY-LE-PENEUX - Panne de courant après un accident sur la D927: Une voiture a percuté un poteau électrique, sur... <a href="https://t.co/86K6937vYo">https://t.co/86K6937vYo</a>	neutre
Une étude qui dérange : 100% d'électricité renouvelable pas plus cher que le nucléaire !   @scoopit <a href="https://t.co/L8KwRBDxmA">https://t.co/L8KwRBDxmA</a>	néгатif
@ComicStrip44 C'est certain. Faut que je me rendorme aussi. Dans 1h, j'ai plus d'électricité et ce jusqu'à 16h...	neutre
je suis en train de câbler, j'ai plus de wifi plus d'électricité plus de téléphone	neutre
Coupure De Courant by Françoise Hardy <a href="https://t.co/Qk4niw8tbl">https://t.co/Qk4niw8tbl</a> #CoupureDeCourant #Fr	neutre
Ca va c'est pas la 5eme panne de courant ☹️	neutre
Probablement pas de stream ce soir parce que SFR refuse de venir réparer la panne si je n'appelle pas, alors qu'ils sont au courant du pb	neutre
@Ma_ow Tu a une startup sudaf qui fait dans les compteur intelligent grâce à la blockchain	neutre
Clumsy, Private Show, Coupure Électrique <a href="https://t.co/0f3Zk7ybEQ">https://t.co/0f3Zk7ybEQ</a>	neutre
Coupure Électrique Just Luv Me Invitation <a href="https://t.co/w2LLkgd3mW">https://t.co/w2LLkgd3mW</a>	neutre
Gaaaars jte dis Et mon tel avait 22% mdrrr <a href="https://t.co/unhCrh4dFe">https://t.co/unhCrh4dFe</a>	neutre
@ntmeminem Coupure électrique	neutre
@alexandreifi C'est le conséquence d'une panne d'alimentation électrique dans le secteur de Pont-l'Évêque plus tôt dans la soirée.	neutre
SNCF. Panne de courant sur la ligne Deauville-Pont-l'Évêque <a href="https://t.co/HeVPK8ieu6">https://t.co/HeVPK8ieu6</a> #train #snf	neutre
Coupure de courant, j'adore j'adhère je valide <a href="https://t.co/xuwXsl32TE">https://t.co/xuwXsl32TE</a>	néгатif
J'entends les rageux qui ont plus d'électricité mdrrr	neutre
tiphaine en pls en message à cause de sa panne de courant, un délire	neutre
La facture EDF qui fache.	néгатif
Les dérives commerciales des concurrents d'EDF pointées du doigt... <a href="https://t.co/eAW1St9Sci">https://t.co/eAW1St9Sci</a>	neutre

Rendez-vous les 14 et 15/6 à Mulhouse pour découvrir nos solutions EDF et @ES_Strasbourg sur le stand et lors d'une... <a href="https://t.co/Rs9Ts0jvLO">https://t.co/Rs9Ts0jvLO</a>	neutre
Une panne électrique dans un datacenter British Airways provoque le chaos à Heathrow La procédure de restaura... <a href="https://t.co/jnaIElAZKM">https://t.co/jnaIElAZKM</a>	négatif
Découvrez comment #EDF utilise le #Numérique au service de la mobilité interne ! #TransfoNum <a href="https://t.co/jnBrgtitaa">https://t.co/jnBrgtitaa</a> <a href="https://t.co/TEt4sFWAQJ">https://t.co/TEt4sFWAQJ</a>	positif
Total participe avec EDF a l elaboration de l usine du futur (efficacité, relations clients)#innovation #industrie40 <a href="https://t.co/pSxzugNC43">https://t.co/pSxzugNC43</a>	positif
À Ivry, les tours EDF reprennent vie <a href="https://t.co/JsIDml3mQl">https://t.co/JsIDml3mQl</a> via @by_beton	positif
Création de la société #Edvance : la refondation du #nucléaire français avance. #edf #areva <a href="https://t.co/leSsViAaY7">https://t.co/leSsViAaY7</a>	positif
Vivement la OuiElectricité <a href="https://t.co/XaspdlWyhs">https://t.co/XaspdlWyhs</a>	positif

## C Extrait du corpus annoté

EDF : pourquoi, malgré son potentiel renouvelable, l'Outre-mer carbure-t-il encore aux énergies fossiles ? - <a href="https://t.co/L52wWvbiyq">https://t.co/L52wWvbiyq</a>	neutre
Propre en RDcongo <a href="https://t.co/HYgniG4gGb">https://t.co/HYgniG4gGb</a>	neutre
Plus d'électricité chez les dupdup super on adore	négatif
♦ Bourse de Paris ♦Le Cac 40 débute la semaine, en hausse Historique, dans les 📊 4.400 📊 points.#Bourse... <a href="https://t.co/o2IX3N1Sl6">https://t.co/o2IX3N1Sl6</a>	neutre
Le @frizoom_fr à la soirée EDF Pulse Hauts-de -France ! On a froid, on est dans le thème ! :- ) <a href="https://t.co/RS9Iyd8kj9">https://t.co/RS9Iyd8kj9</a>	neutre
Sans ces équipements électriques d'EDF, pas de #MacBookPro ! <a href="https://t.co/Qg6bjf3Esa">https://t.co/Qg6bjf3Esa</a>	neutre
Jviens de survivre à une coupure de courant d'environ 30 min jpeux vous dire que jme plaindrai pus jamais du bruit qu'il fait	négatif
.@iot_business5 startups du moment dans la #SmartCity .@veniamworks .@Plume_Labs .@enevo #IoT #smartgrid <a href="https://t.co/gey7C0ebOF">https://t.co/gey7C0ebOF</a>	neutre
@NereaBP_ déjà ouvre pas ton frigo/congel pour rien ! Tu as contact edf urgence ? Ou la mairie savoir si un plan a été déclenché ?	neutre
#QAG #Fessenheim @jeandesessard En contrepartie de la fermeture de Fessenheim, #EDF recevra la somme excessive de 490 M€ 1/3	négatif
@TotoCelclubow Si ça se trouve ils se reconvertissent en simples agents EDF	neutre
[#Obsdelaville] Point à date ac @cityofsound sur la ville du XXIe siècle & son urbanisme en réseau #smartGrid <a href="https://t.co/DKqzkZ65fp">https://t.co/DKqzkZ65fp</a>	neutre
New Marriage Proposal DRBM37880 =&gt;Sex : Male, Age : 36, Nati.. <a href="https://t.co/ztRX8TwtCY">https://t.co/ztRX8TwtCY</a>	neutre
J'aime une vidéo @YouTube : "La vidéo que EDF veut interdire ! ÉLECTRICITÉ - GRATUIT A VIE..." à l'adresse	négatif
#Soldes2017 Philips Lot de 2 Ampoules LED Spot Culot GU10 5W Consommés Équivalent 50W Partenariat Philips/EDF 📢 <a href="https://t.co/oKERchYnEb">https://t.co/oKERchYnEb</a>	neutre
Areva NP: protocole d'accord entre EDF et MHI pour une possible entrée du japonais au capital: EDF a signé ma... <a href="https://t.co/r1bxKrql4y">https://t.co/r1bxKrql4y</a>	neutre
Quand il y a sowee obligé * <a href="https://t.co/k7rOBU7PDI">https://t.co/k7rOBU7PDI</a>	neutre
#ÉNERGIE EDF: nouveau revers pour le projet nucléaire Hinkley Point <a href="https://t.co/iamxjji4hd">https://t.co/iamxjji4hd</a>   <a href="https://t.co/qjLRw5Kyr2">https://t.co/qjLRw5Kyr2</a>	négatif
Au coeur du #barrage #EDF de #Castillon pour l'émission #ÉchappéesBelles @So_Jovillard <a href="https://t.co/2KTBF4tjho">https://t.co/2KTBF4tjho</a>	neutre
News - Le nucléaire ne survit plus que par la triche: La décision prise par le conseil d'administration d'EDF... <a href="https://t.co/8UiB4h2nco">https://t.co/8UiB4h2nco</a>	négatif
Une personne électrosensible obtient le retrait de son compteur « intelligent » <a href="https://t.co/KD7L5cFd3l">https://t.co/KD7L5cFd3l</a> #ecologie #feedly	neutre
"Nucléaire : EDF va arrêter cinq réacteurs dont un à Gravelines pour audit" <a href="https://t.co/RL6LZ58n3o">https://t.co/RL6LZ58n3o</a> #actualites #feedly	neutre
coupure de courant j'adore	négatif
Ptn d'merde y'a toujours pas d'électricité dans mon bled	négatif
@GilH94 Bjr, je suis navrée de cette situation. Afin de vous aider, donnez-moi votre n° de client en DM. Merci. Pascale	neutre
Centrale de Hinkley Point : "Ça peut mettre EDF en péril"	négatif

<a href="https://t.co/4eraEMANg6">https://t.co/4eraEMANg6</a>	
Stagiaire data-analyse wind en productie: EDF	neutre
Luminus <a href="http://stst.tw/a1438722">http://stst.tw/a1438722</a> #bejobs #IT	
Après 3h de coupure de courant j'ai enfin de la wifi !	neutre
Dites Non au compteur #Linky ! <a href="https://t.co/AvXt97wNLN">https://t.co/AvXt97wNLN</a> via @planetes360	négatif
@LeMetro EDF passer sur les indiens Mapuches au Chili.Rejet généralisé par le terminale de GNL <a href="https://t.co/431ceA9nPj">https://t.co/431ceA9nPj</a>	neutre
@DuoMortel Merci à vous également bonne journée.	neutre
@Francedz213 @edf_client @villedemassy un coup des grévistes..	neutre
Putain les gars j'ai plus d'électricité c'est man vs wild <a href="https://t.co/XKJ0yuFAtA">https://t.co/XKJ0yuFAtA</a>	négatif
@EmilieLacaille Je reste à votre disposition en cas de besoin et je vous souhaite une bonne journée. Matthieu	neutre
Retweeted 3ndoffline (@3ndoffline): <a href="https://t.co/l3cwteCelz">https://t.co/l3cwteCelz</a> ça bouge ! 308 communes refusent cette saloperie de co... <a href="https://t.co/l3cwteCelz">https://t.co/l3cwteCelz</a>	négatif
Information importante où les #norme ouvrent des marchés <a href="https://t.co/7Hi7ushldq">https://t.co/7Hi7ushldq</a>	neutre
cyberaction #linky #enidis #électricité Linky : Soutenons maires attaqués par ENEDIS <a href="https://t.co/ja5KVg6o93">https://t.co/ja5KVg6o93</a> via @cyberacteurs	négatif
Très bien je rentre à l'appart et j'ai plus d'électricité je suis ravie	négatif
@lemondefr Hollande se trouve aux chiliennes disent qu'ils soutiennent les anergies renouvelables et derrière EDF c... <a href="https://t.co/O158CURDlx">https://t.co/O158CURDlx</a>	neutre
@Forum_Ecolo Et oui, #Linky, on l'aura tous dans le cul de la même façon !	négatif
EDF & MOI android app is now available at lilouw85 repository. <a href="https://t.co/pDqRcjvUFh">https://t.co/pDqRcjvUFh</a>	neutre
@tomtomson Bonjour, le courant est-il revenu ? Matthieu	neutre
@Foutezy Nous essayons de vous répondre par message direct. Merci de nous suivre sur Twitter pour que nous puissions vous l'envoyer.	neutre
L'AMF perquisitionne le siège d'EDF, en cause la communication financière depuis 2013 notamment sur le #nucléaire. <a href="https://t.co/05aiVqDPfu">https://t.co/05aiVqDPfu</a>	neutre
#ALaUne Hinkley Point : le comité central d'entreprise d'EDF débouté en référé: Le comité central d'entrepris... <a href="https://t.co/maYshsW0nw">https://t.co/maYshsW0nw</a>	négatif
Important bug chez EDF : un million de clients prélevés deux fois <a href="https://t.co/elljhg64Yu">https://t.co/elljhg64Yu</a> <a href="https://t.co/amaGLtLTHw">https://t.co/amaGLtLTHw</a>	négatif
@GossipRoomOff c'est pas une coupure de courant qui va changer leur vie	neutre
@_Cristo_P Bjr, concernant la mesure de nationalisation d'#EDF, je vous invite à consulter notre décryptage → <a href="https://t.co/xb4QcMckJy">https://t.co/xb4QcMckJy</a> , cdlt	neutre
0000 Coupure de courant 🤔	neutre
#énergie @RonanDantec Notre problème principal est effondrement du prix de gros passé de 42€ le MW à 28€, #EDF estimant limite viable à 37€	neutre

EDF : pourquoi, malgré son potentiel renouvelable, l'Outre-mer carbure-t-il encore aux énergies fossiles ? - <a href="https://t.co/L52wWvbiyq">https://t.co/L52wWvbiyq</a>	neutre
EDF dégaîne Sowe, un hub pour piloter la smart home et une marque IoT <a href="https://t.co/G1XRySieHU">https://t.co/G1XRySieHU</a> <a href="https://t.co/jg0brutzCB">https://t.co/jg0brutzCB</a>	positif
Profitez des vacances pour découvrir le #belvdère #EDF de #SainteTulle Infos pratiques : <a href="https://t.co/7IVxFvMBFc">https://t.co/7IVxFvMBFc</a> <a href="https://t.co/oGgQGMErqD">https://t.co/oGgQGMErqD</a>	positif
Kris Vervaeet explique pourquoi l'accompagnement d'EDF est fondamental po... <a href="https://t.co/DGBGDyXGJl">https://t.co/DGBGDyXGJl</a> via @YouTube	positif
Le @frizoom_fr à la soirée EDF Pulse Hauts-de -France ! On a froid, on est dans le thème ! :- ) <a href="https://t.co/RS9lyd8kj9">https://t.co/RS9lyd8kj9</a>	neutre
Sans ces équipements électriques d'EDF, pas de #MacBookPro ! <a href="https://t.co/Qg6bjf3Esa">https://t.co/Qg6bjf3Esa</a>	neutre
Jviens de survivre à une coupure de courant d'environ 30 min jpeux vous dire que jme plaindrai pus jamais du bruit qu'il fait	neutre
.@iot_business5 startups du moment dans la #SmartCity .@veniamworks .@Plume_Labs .@enevo #IoT #smartgrid <a href="https://t.co/gey7C0ebOF">https://t.co/gey7C0ebOF</a>	positif
@NereaBP_ déjà ouvre pas ton frigo/congel pour rien ! Tu as contact edf urgence ? Ou la mairie savoir si un plan a été déclenché ?	neutre
#QAG #Fessenheim @jeandesessard En contrepartie de la fermeture de Fessenheim, #EDF recevra la somme excessive de 490 M€ 1/3	neutre
@TotoCelplubow Si ça se trouve ils se reconvertissent en simples agents EDF	neutre
[#Obsdelaville] Point à date ac @cityofsound sur la ville du XXIe siècle & son urbanisme en réseau #smartGrid <a href="https://t.co/DKqzkZ65fp">https://t.co/DKqzkZ65fp</a>	neutre
New Marriage Proposal DRBM37880 =>Sex : Male, Age : 36, Nati.. <a href="https://t.co/ztRX8TwtCY">https://t.co/ztRX8TwtCY</a>	neutre
J'aime une vidéo @YouTube : "La vidéo que EDF veut interdire ! ÉLECTRICITÉ - GRATUIT A VIE..." à l'adresse	neutre
#Soldes2017 Philips Lot de 2 Ampoules LED Spot Culot GU10 5W Consommés Équivalent 50W Partenariat Philips/EDF  <a href="https://t.co/oKERchYnEb">https://t.co/oKERchYnEb</a>	neutre
Areva NP: protocole d'accord entre EDF et MHI pour une possible entrée du japonais au capital: EDF a signé ma... <a href="https://t.co/r1bxKrql4y">https://t.co/r1bxKrql4y</a>	neutre
Quand il y a sowe obligé ☆ <a href="https://t.co/k7rOBU7PDI">https://t.co/k7rOBU7PDI</a>	neutre
#ÉNERGIE EDF: nouveau revers pour le projet nucléaire Hinkley Point <a href="https://t.co/iamxjji4hd">https://t.co/iamxjji4hd</a>   <a href="https://t.co/qjLRw5Kyr2">https://t.co/qjLRw5Kyr2</a>	neutre
Au coeur du #barrage #EDF de #Castillon pour l'émission #ÉchappéesBelles @So_Jovillard <a href="https://t.co/2KTBF4tjho">https://t.co/2KTBF4tjho</a>	neutre
News - Le nucléaire ne survit plus que par la triche: La décision prise par le conseil d'administration d'EDF... <a href="https://t.co/8UiB4h2nco">https://t.co/8UiB4h2nco</a>	neutre
Une personne électrosensible obtient le retrait de son compteur « intelligent » <a href="https://t.co/KD7L5cFd3I">https://t.co/KD7L5cFd3I</a> #ecologie #feedly	neutre
"Nucléaire : EDF va arrêter cinq réacteurs dont un à Gravelines pour audit" <a href="https://t.co/RL6LZ58n3o">https://t.co/RL6LZ58n3o</a> #actualites #feedly	neutre
coupure de courant j'adore	neutre
Ptn d'merde y'a toujours pas d'électricité dans mon bled	neutre
@GilH94 Bjr, je suis navrée de cette situation. Afin de vous aider,	neutre

donnez-moi votre n° de client en DM. Merci. Pascale	
Centrale de Hinkley Point : "Ça peut mettre EDF en péril" <a href="https://t.co/4eraEMANg6">https://t.co/4eraEMANg6</a>	neutre
Stagiaire data-analyse wind en productie: EDF Luminus <a href="http://stst.tw/a1438722">http://stst.tw/a1438722</a> #bejobs #IT	neutre
Après 3h de coupure de courant j'ai enfin de la wifi !	neutre
Dites Non au compteur #Linky ! <a href="https://t.co/AvXt97wNLN">https://t.co/AvXt97wNLN</a> via @planetes360	neutre
@LeMetro EDF passer sur les indiens Mapuches au Chili.Rejet généralisé par le terminale de GNL <a href="https://t.co/431ceA9nPj">https://t.co/431ceA9nPj</a>	neutre
@DuoMortel Merci à vous également bonne journée.	positif
@Francedz213 @edf_client @villedemassy un coup des grévistes..	neutre
Putain les gars j'ai plus d'électricité c'est man vs wild <a href="https://t.co/XKJ0yuFAtA">https://t.co/XKJ0yuFAtA</a>	neutre
@EmilieLacaille Je reste à votre disposition en cas de besoin et je vous souhaite une bonne journée. Matthieu	neutre
Retweeted 3ndofline (@3ndofline): <a href="https://t.co/l3cwteCelz">https://t.co/l3cwteCelz</a> ça bouge ! 308 communes refusent cette saloperie de co... <a href="https://t.co/l3cwteCelz">https://t.co/l3cwteCelz</a>	neutre
Information importante où les #norme ouvrent des marchés <a href="https://t.co/7Hi7ushldq">https://t.co/7Hi7ushldq</a>	neutre
cyberaction #linky #enidis #électricité Linky : Soutenons maires attaqués par ENEDIS <a href="https://t.co/ja5KVg6o93">https://t.co/ja5KVg6o93</a> via @cyberacteurs	neutre
Très bien je rentre à l'appart et j'ai plus d'électricité je suis ravie	neutre
@lemondefr Hollande se trouve aux chiliennes disent qu'ils soutiennent les anergies renouvelables et derrière EDF c... <a href="https://t.co/O158CURDlx">https://t.co/O158CURDlx</a>	neutre
@Forum_Ecolo Et oui, #Linky, on l'aura tous dans le cul de la même façon !	neutre
EDF & MOI android app is now available at lilouw85 repository. <a href="https://t.co/pDqRcJvUFh">https://t.co/pDqRcJvUFh</a>	neutre
@tomtomson Bonjour, le courant est-il revenu ? Matthieu	neutre
@Foutezy Nous essayons de vous répondre par message direct. Merci de nous suivre sur Twitter pour que nous puissions vous l'envoyer.	neutre
L'AMF perquisitionne le siège d'EDF, en cause la communication financière depuis 2013 notamment sur le #nucléaire. <a href="https://t.co/05aiVqDPfu">https://t.co/05aiVqDPfu</a>	neutre
#ALaUne Hinkley Point : le comité central d'entreprise d'EDF débouté en référé: Le comité central d'entrepris... <a href="https://t.co/maYshsW0nw">https://t.co/maYshsW0nw</a>	neutre
Important bug chez EDF : un million de clients prélevés deux fois <a href="https://t.co/elljhg64Yu">https://t.co/elljhg64Yu</a> <a href="https://t.co/amaGLtLTHw">https://t.co/amaGLtLTHw</a>	neutre
@GossipRoomOff c'est pas une coupure de courant qui va changer leur vie	neutre
@_Cristo_P Bjr, concernant la mesure de nationalisation d'#EDF, je vous invite à consulter notre décryptage → <a href="https://t.co/xb4QcMcKJy">https://t.co/xb4QcMcKJy</a> , cdl	neutre
0000 Coupure de courant 🤔	neutre
#énergie @RonanDantec Notre problème principal est effondrement du prix de gros passé de 42€ le MW à 28€, #EDF estimant limite viable à 37€	neutre

## D Liste des traits et relation renvoyés par XIP

Liste de tous les éléments renvoyés par XIP via l'API Python pour la phrase :  
*Bon une coupure de courant c'est marrant 5 min mais là je commence à me faire chier*

```
[ 'Bon', 'bon_ADJ', 'bon_[QUEP:+', 'bon_POURSVINF:+',
'bon_POURSN:+', 'bon_ENVERSSN:+', 'bon_ENSN:+',
'bon_DESVINF:+', 'bon_DANSSN:+', 'bon_CONTRESN:+',
'bon_AVECSN:+', 'bon_ASVINF:+', 'bon_ASN:+',
'bon_!SVINFDIR:+', 'bon_IMPERSO:+', 'bon_STARTBIS:+',
'bon_MAJ:+', 'bon_SFPOUR:+', 'bon_SFENVERS:+',
'bon_SFEN:+', 'bon_SFDANS:+', 'bon_SFCONTRE:+',
'bon_SFAVEC:+', 'bon_SFA:+', 'bon_MASC:+', 'bon_SG:+',
'bon_ADJ:+', 'bon_ADJ2:+', 'bon_!START:+',
'bon_LAST:+', 'bon_FIRST:+]', 'bon_[QUEP:+,
POURSVINF:+,POURSN:+,ENVERSSN:+,ENSN:+,DESVINF:+,
DANSSN:+,CONTRESN:+,AVECSN:+,ASVINF:+,ASN:+,
!SVINFDIR:+,IMPERSO:+,STARTBIS:+,MAJ:+,SFPOUR:+,
SFENVERS:+,SFEN:+,SFDANS:+,SFCONTRE:+,SFAVEC:+,
SFA:+,MASC:+,SG:+,ADJ:+,ADJ2:+,!START:+,LAST:+,
FIRST:+]', 'une', 'un_DET', 'un_[FONC:+',
'un_CLOSED:+', 'un_FEM:+', 'un_SG:+', 'un_INDEF:+',
'un_DET:+', 'un_FIRST:+]', 'un_[FONC:+,CLOSED:+,
FEM:+,SG:+,INDEF:+,DET:+,FIRST:+]', 'coupure',
'coupure_NOUN', 'coupure_[!PARSN:+', 'coupure_!ETSN:+',
'coupure_!ENTRESN:+', 'coupure_!DESN:+',
'coupure_!ASN:+', 'coupure_SFPAR:+',
'coupure_SFENTRE:+', 'coupure_SFDE:+', 'coupure_SFA:+',
'coupure_FEM:+', 'coupure_SG:+', 'coupure_P3:+',
'coupure_!NOUN:+]', 'coupure_[!DESN:+',
'coupure_!AVECSN:+', 'coupure_SFDE:+',
'coupure_SFAVEC:+', 'coupure_FEM:+', 'coupure_SG:+',
'coupure_P3:+', 'coupure_!NOUN:+]', 'coupure_[!PARSN:+,
!ETSN:+,!ENTRESN:+,!DESN:+,!ASN:+,SFPAR:+,SFENTRE:+,
SFDE:+,SFA:+,FEM:+,SG:+,P3:+,!NOUN:+]', 'de',
'de_PREP', 'de_[PREPINF:+', 'de_SFDE:+',
'de_FORM:FDE', 'de_FONC:+', 'de_PREP:+',
'de_DIR:+', 'de_FIRST:+]', 'de_[PREPINF:+,
SFDE:+,FORM:FDE,FONC:+,PREP:+,DIR:+,FIRST:+]',
'courant', 'courant_NOUN', 'courant_[DESN:+',
'courant_SFDE:+', 'courant_FONC:+', 'courant_MASC:+',
'courant_SG:+', 'courant_P3:+', 'courant_NOUN:+',
'courant_LAST:+]', 'courant_[DESN:+,SFDE:+,FONC:+,MASC:+,
SG:+,P3:+,NOUN:+,LAST:+]', "c'", 'ce_PRON', 'ce_[CLOSED:+',
'ce_MASC:+', 'ce_SG:+', 'ce_DEM:+', 'ce_PRON:+', 'ce_FIRST:+]',
'ce_[CLOSED:+,MASC:+,SG:+,DEM:+,PRON:+,FIRST:+]', 'est',
'\xc3\xaatre_VERB', '\xc3\xaatre_[QUEP:+',
'\xc3\xaatre_POURSN:+', '\xc3\xaatre_LOCSN:+',
'\xc3\xaatre_IL:+', '\xc3\xaatre_EN:+',
```

'\xc3\xaatre\_DESVINF:+', '\xc3\xaatre\_DESN:+',  
 '\xc3\xaatre\_CONTRESN:+', '\xc3\xaatre\_AVOIR:+',  
 '\xc3\xaatre\_ASVINF:+', '\xc3\xaatre\_ASN:+',  
 '\xc3\xaatre\_SVINFDIR:+', '\xc3\xaatre\_SN:+',  
 '\xc3\xaatre\_SADJDESVINF:+', '\xc3\xaatre\_SADJ1:+',  
 '\xc3\xaatre\_SADJ:+', '\xc3\xaatre\_SFPOUR:+',  
 '\xc3\xaatre\_SFDE:+', '\xc3\xaatre\_SFCONTRE:+',  
 '\xc3\xaatre\_SFA:+', '\xc3\xaatre\_FORM:FETRE',  
 '\xc3\xaatre\_FONC:+', '\xc3\xaatre\_SG:+',  
 '\xc3\xaatre\_P3:+', '\xc3\xaatre\_PRE:+',  
 '\xc3\xaatre\_IND:+', '\xc3\xaatre\_AUX:+',  
 '\xc3\xaatre\_COPULE:+', '\xc3\xaatre\_VERB:+',  
 '\xc3\xaatre\_LAST:+]',  
 '\xc3\xaatre\_[QUEP:+,POURSN:+,LOCSN:+,IL:+,EN:+,DESVINF:+,  
 DESN:+,CONTRESN:+,AVOIR:+,ASVINF:+,ASN:+,SVINFDIR:+,SN:+,  
 SADJDESVINF:+,SADJ1:+,SADJ:+,SFPOUR:+,SFDE:+,SFCONTRE:+,  
 SFA:+,FORM:FETRE,FONC:+,SG:+,P3:+,PRE:+,IND:+,AUX:+,  
 COPULE:+,VERB:+,LAST:+]', 'marrant', 'marrant\_ADJ',  
 'marrant\_[DESVINF:+', 'marrant\_!SVINFDIR:+',  
 'marrant\_IMPERSO:+', 'marrant\_FONC:+', 'marrant\_MASC:+',  
 'marrant\_SG:+', 'marrant\_ADJ:+', 'marrant\_FIRST:+]',  
 'marrant\_[DESVINF:+,!SVINFDIR:+,IMPERSO:+,FONC:+,  
 MASC:+,SG:+,ADJ:+,FIRST:+]', '5', '5\_NUM', '5\_[FONC:+',  
 '5\_!CLOSED:+', '5\_NUM:+]', '5\_[FONC:+,!CLOSED:+,NUM:+]',  
 'min', 'min\_NOUN', 'min\_[FONC:+', 'min\_CLOSED:+',  
 'min\_MASC:+', 'min\_FEM:+', 'min\_PL:+', 'min\_SG:+',  
 'min\_P3:+', 'min\_MEASURE:+', 'min\_ABR:+', 'min\_NOUN:+',  
 'min\_LAST:+]', 'min\_[FONC:+,CLOSED:+,MASC:+,FEM:+,  
 PL:+,SG:+,P3:+,MEASURE:+,ABR:+,NOUN:+,LAST:+]', ',',  
 ',\_PUNCT', ',\_[FORM:FCM', ',\_PUNCT:+]', ',\_[FORM:FCM,  
 PUNCT:+]', 'mais', 'mais\_COORD', 'mais\_[COORD:+',  
 'mais\_FIRST:+]', 'mais\_[COORD:+,FIRST:+]', 'la',  
 'le\_PRON', 'le\_[!CLOSED:+', 'le\_FEM:+', 'le\_SG:+',  
 'le\_P3:+', 'le\_CLIT:+', 'le\_ACC:+', 'le\_!PRON:+',  
 'le\_LAST:+', 'le\_FIRST:+]', 'le\_[!CLOSED:+,FEM:+,  
 SG:+,P3:+,CLIT:+,ACC:+,!PRON:+,LAST:+,FIRST:+]',  
 'je', 'je\_PRON', 'je\_[FONC:+', 'je\_CLOSED:+',  
 'je\_MASC:+', 'je\_FEM:+', 'je\_SG:+', 'je\_P1:+',  
 'je\_CLIT:+', 'je\_NOM:+', 'je\_PRON:+', 'je\_FIRST:+]',  
 'je\_[FONC:+,CLOSED:+,MASC:+,FEM:+,SG:+,P1:+,CLIT:+,  
 NOM:+,PRON:+,FIRST:+]', 'commence', 'commencer\_VERB',  
 'commencer\_[!PARSVINF:+', 'commencer\_!PARSN:+',  
 'commencer\_!DESVINF:+', 'commencer\_!AVOIR:+',  
 'commencer\_!ASVINF:+', 'commencer\_!SVINFDIR:+',  
 'commencer\_!SN:+', 'commencer\_SFPAR:+', 'commencer\_SG:+',  
 'commencer\_P1:+', 'commencer\_PRE:+', 'commencer\_IND:+',  
 'commencer\_VERB:+]', 'commencer\_[!PARSVINF:+',  
 'commencer\_!PARSN:+', 'commencer\_!DESVINF:+',  
 'commencer\_!AVOIR:+', 'commencer\_!ASVINF:+',  
 'commencer\_!SVINFDIR:+', 'commencer\_!SN:+',

'commencer\_SFPAR:+', 'commencer\_SG:+', 'commencer\_P1:+',  
 'commencer\_PRE:+', 'commencer\_SUBJ:+',  
 'commencer\_VERB:+', 'commencer\_ [!PARSVINF:+,  
 !PARSN:+,!DESVINF:+,!AVOIR:+,!ASVINF:+,!SVINFDIR:+,  
 !SN:+,SFPAR:+,SG:+,P1:+,PRE:+,IND:+,VERB:+] ',  
 '\xc3\xa0', '\xc3\xa0\_NOUN', '\xc3\xa0\_ [FONC:+',  
 '\xc3\xa0\_GUESSED:+', '\xc3\xa0\_MASC:+', '\xc3\xa0\_FEM:+',  
 '\xc3\xa0\_PL:+', '\xc3\xa0\_SG:+', '\xc3\xa0\_P3:+',  
 '\xc3\xa0\_NOUN:+', '\xc3\xa0\_LAST:+', '\xc3\xa0\_FIRST:+] ',  
 '\xc3\xa0\_ [FONC:+,GUESSED:+,MASC:+,FEM:+,PL:+,SG:+,  
 P3:+,NOUN:+,LAST:+,FIRST:+] ', 'me', 'me\_PRON',  
 'me\_ [FONC: FIOBJ', 'me\_CLOSED:+', 'me\_MASC:+',  
 'me\_FEM:+', 'me\_SG:+', 'me\_P1:+', 'me\_CLIT:+',  
 'me\_DAT:+', 'me\_GEN:+', 'me\_ACC:+', 'me\_PRON:+',  
 'me\_FIRST:+] ', 'me\_ [FONC: FIOBJ, CLOSED:+, MASC:+,  
 FEM:+, SG:+, P1:+, CLIT:+, DAT:+, GEN:+, ACC:+, PRON:+,  
 FIRST:+] ', 'faire', 'faire\_VERB', 'faire\_ [SEN:+',  
 'faire\_ILSE:+', 'faire\_SE:+', 'faire\_QUEP:+',  
 'faire\_POURS\_N:+', 'faire\_DESN2:+', 'faire\_AVOIR:+',  
 'faire\_ASN:+', 'faire\_SVINFDIR:+', 'faire\_SVINF:+',  
 'faire\_SN:+', 'faire\_SADJ:+', 'faire\_INF\_COD:+',  
 'faire\_REFLEXTYPE:+', 'faire\_REFLEXIVE:+',  
 'faire\_SFPOUR:+', 'faire\_SFDE:+', 'faire\_SFA:+',  
 'faire\_FORM: FFAIRE', 'faire\_FONC:+', 'faire\_INF:+',  
 'faire\_VERB:+', 'faire\_LAST:+] ', 'faire\_ [SEN:+,  
 ILSE:+, SE:+, QUEP:+, POURS\_N:+, DESN2:+, AVOIR:+, ASN:+,  
 SVINFDIR:+, SVINF:+, SN:+, SADJ:+, INF\_COD:+, REFLEXTYPE:+,  
 REFLEXIVE:+, SFPOUR:+, SFDE:+, SFA:+, FORM: FFAIRE,  
 FONC:+, INF:+, VERB:+, LAST:+] ', 'chier',  
 'chier\_VERB', 'chier\_ [EN:+', 'chier\_AVOIR:+',  
 'chier\_FONC:+', 'chier\_INF:+', 'chier\_VERB:+',  
 'chier\_LAST:+', 'chier\_FIRST:+] ', 'chier\_ [EN:+,  
 AVOIR:+, FONC:+, INF:+, VERB:+, LAST:+, FIRST:+] ',  
 '5\_MOD-N\_min', 'me\_CPL-V\_faire',  
 'coupure\_SUJ-V\_\xc3\xaatre', 'je\_SUJ-V\_commencer',  
 '\xc3\xa0\_COD-V\_commencer', 'chier\_COD-V\_faire',  
 'faire\_COD-V\_commencer', 'courant\_MOD-N\_coupure',  
 'marrant\_MOD-N\_min', 'min\_ATB-SO\_\xc3\xaatre',  
 'commencer\_PRECOMMA', 'coupure\_CLOSEDNP',  
 'ce\_CLOSEDNP', '\_CLOSEDNP', 'je\_CLOSEDNP']

## E Termes les plus spécifiques

Ci-dessous, nous avons rapporté les termes les plus spécifiques à chaque corpus, le corpus négatif. Comme indiqué précédemment, ces termes ont été extraits à l'aide du logiciel TXM.

### E.1 Corpus positif

- innovation
- merci
- bravo
- smartgrid
- open
- associer
- découvrir
- pulse
- concours
- contribuer
- accompagner
- fier
- startup
- transition énergétique
- enedis
- sosee
- énergie

### E.2 Corpus négatif

- courant
- coupure
- merde
- plus
- putain
- contre
- électricité
- vie
- non
- augmenter
- augmenter
- bloquer
- LINKY
- Greenpeace
- payer
- adorer
- sérieux
- super
- CGT
- Mauvais

## F Résultats détaillés

Nous présentons ci-dessous les résultats détaillés de tous les algorithmes testés pour le calcul de la baseline.

Dans l'optique de la construction de deux modèles différents pour chaque polarité (un modèle pour déterminer les tweets négatifs ou neutres et un pour les tweets positifs ou neutres), les résultats ont été testés sur le corpus positif et sur le corpus négatif séparément, les résultats des deux tests sont indiqués.

### F.1 Régression logistique

NEGATIF	Précision	Rappel	F-mesure
Neutre	0.70	0.73	0.72
Négatif	0.74	0.71	0.73
Total	0.72	0.72	0.72

POSITIF	Précision	Rappel	F-mesure
Neutre	0.74	0.73	0.73
Négatif	0.70	0.71	0.70
Total	0.72	0.72	0.72

### F.2 Bayes Multinomial

NEGATIF	Précision	Rappel	F-mesure
Neutre	0.69	0.52	0.59
Négatif	0.64	0.79	0.71
Total	0.67	0.66	0.65

POSITIF	Précision	Rappel	F-mesure
Neutre	0.74	0.72	0.73
Négatif	0.69	0.72	0.71
Total	0.72	0.72	0.72

### F.3 Bayes Gaussien

NEGATIF	Précision	Rappel	F-mesure
Neutre	0.70	0.47	0.56
Négatif	0.62	0.81	0.70
Total	0.66	0.65	0.64

POSITIF	Précision	Rappel	F-mesure
Neutre	0.70	0.55	0.62
Négatif	0.59	0.73	0.65
Total	0.65	0.64	0.63

### F.4 K plus proches voisins (n=3)

NEGATIF	Précision	Rappel	F-mesure
Neutre	0.54	0.69	0.61
Négatif	0.62	0.46	0.53
Total	0.58	0.57	0.57

POSITIF	Précision	Rappel	F-mesure
Neutre	0.60	0.88	0.71
Négatif	0.71	0.34	0.46
Total	0.65	0.62	0.59

### F.5 Arbres de décision

NEGATIF	Précision	Rappel	F-mesure
Neutre	0.64	0.69	0.66
Négatif	0.69	0.64	0.67
Total	0.67	0.67	0.67

POSITIF	Précision	Rappel	F-mesure
Neutre	0.65	0.70	0.67
Négatif	0.63	0.58	0.60
Total	0.64	0.64	0.64

### F.6 SVM

NEGATIF	Précision	Rappel	F-mesure
Neutre	0.71	0.70	0.70
Négatif	0.72	0.74	0.73
Total	0.72	0.72	0.72

POSITIF	Précision	Rappel	F-mesure
Neutre	0.73	0.72	0.72
Négatif	0.69	0.69	0.69
Total	0.71	0.71	0.71