Tenter la *Fake News Challenge*

Une proposition d'utilisation de la *Stance Detection* pour mieux anticiper le phénomène des *Fake News*

UNIVERSITÉ DE GENÈVE

FACULTÉ DES LETTRES

Enzo Poggio

Superviseuse: Pr. Paola Merlo

Faculté des Lettres
Université de Genève

Ce mémoire est l'achèvement du
Master d'Informatique pour Sciences Humaines

Université de Genève Mai 2018

Déclaration

Ce travail original est le fruit de ma réflexion. Je déclare n'avoir jamais utilisé son contenu pour un autre travail ou dans une autre université. Son contenu est inspiré et sourcé de nombreuses références. Ce travail contient environ 180.000 signes, une cinquantaine de pages, 37 tableaux et 15 figures.

Enzo Poggio
Mai 2018

Remerciements

J'aimerais remercier en premier lieu l'Université de Genève, ma superviseuse, la Professeure Paola Merlo et les membres de mon jury.

J'aimerais remercier pour leurs nombreuses relectures, corrections et conseils ma mère, Édith Certain, mon ami, David Burkhard et mon amie, Alice Badel.

Pour leurs précieux conseils sur lateX, en math et en rédaction, je remercie mes amis Allan Fries, Miguel Van Vlasselaer, Sahar Al Jalbout et Timothée Premat.

Enfin pour m'avoir aidé à garder le moral tout au long de la rédaction je remercie mes amis Delphine Vidal et Léandre Phillippon.

Pour m'avoir aidé à décompresser, je remercie toutes les personnes du Centre Saint-Boniface.

Et pour les nombreuses inspirations sceptiques, je souhaite remercier les propriétaires des chaînes Youtube suivantes : Aude WTFake, Esprit Critique, Hygiène Mental, Instant Sceptique, La statistique expliquée à mon chat, La Tronche en Biais, Le Pharmacien, Mycéliums, Officiel DEFAKATOR, Sciences4All, Un Monde Riant et Les Shadoks : Shadok Tube.

Résumé

Dans ce mémoire nous apportons des réponses à ces questions : y a-t-il une bonne définition des fausses nouvelles (*Fake News*) ? Peut-on parler de *Fake News* de manière objective ? Qu'implique le terme *Fake News* ? D'ou proviennent ces fameuses *Fake News* et que font-elles ? Et pourquoi est-ce important d'en parler ?

Puis nous présentons un bref état actuel de la *Stance Detection*. Nous montrons comment elle peut être utile pour détecter les *Fake News* en présentant deux tâches, à savoir la *SemEval-2016 Task 6* et la *Fake News Challenge*.

Enfin nous présentons et discutons nos propres résultats à la *Fake News Challenge* (dont certains dépassent l'état de l'art) et notre utilisation des combinaisons de modèles et de l'apprentissage par ensemble.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction : Qu'est-ce qu'une *Fake News* ?

Lors des vœux présidentiels annuels de 2018 en France, « Emmanuel Macron s'est trouvé un ennemi commun avec les médias : la lutte contre les fausses nouvelles, ou *Fake News* comme le disent les Anglo-saxons » [9]. Le but du président est de sanctionner juridiquement les fausses nouvelles et les maux qu'elles engendrent. Le but second est de protéger la vie démocratique, surtout en période d'élections. Donner une définition légale aux fausses nouvelles nous interroge sur leur nature. Pour le sens commun ou dans l'usage courant, la fausse nouvelle est « tout ce avec quoi je ne suis pas d'accord » . On voit tout de suite les limites de ce type de définition. Deux personnes avec des opinions contraires traiteraient les informations de l'autre comme des fausses nouvelles. Alors comment faut-il faire ? Y a-t-il un bonne définition des fausses nouvelles (*Fake News*) ? Peut-on parler de *Fake News* de manière objective ? Qu'implique le terme *Fake News* ? D'ou proviennent ces fameuses *Fake News* et que font-elles ? Et pourquoi est-ce important d'en parler ?

Nous répondrons à ces questions en présentant dans cette partie notre définition des *Fake News*, leur raison d'être, leur provenance, leur moyens de propagation et les risques qu'elles entraînent. Puis enfin nous présenterons un moyen pour les détecter.

1.1 Vers une définition du concept de *Fake News*

1.1.1 Produire des *Fake News* ; une intention de tromper

Un énoncé peut être vrai ou faux. Il est vrai s'il est en adéquation avec la réalité telle que je la perçois. Il est faux s'il ne correspond pas à la réalité telle que je la perçois. Les nouvelles (*news*) sont des énoncés. Est-ce que l'énoncé « J'aime les tartes » est une *news* ? Non, toutes les *news* sont des énoncés ; mais seuls certains énoncés sont des *news*. Des journalistes ont fait une liste non exhaustive des critères d'une *news* (Tony Harcup et Deirdre O'Neill[16]). La *news* est une histoire d'une personne de pouvoir ou célèbre. *« Concert annulé de Bertrand Cantat [...] »* [1] Elle est parfois une bonne ou une mauvaise nouvelle. *« Johnny Hallyday est mort à l'âge de 74 ans »* [4]. Souvent elle est une surprise ou un phénomène qui concerne beaucoup de personnes. « *Ouragan Maria. L'état de catastrophe naturelle* [déclaré ...] » [29]. Certaines sont juste des suivis. *« Accusations contre Médiapart : Edwy Plenel répond à Nicolas Sarkozy »* [6]. Et parfois ce sont juste des divertissements ou des faits divers. *« Alexia Mori enceinte : Elle dévoile combien de kilos elle a pris ! »* [33]. En somme, retenons qu'une nouvelle est une histoire en lien avec la réalité. Ce qui nous intéresse, c'est sa véracité selon notre interprétation du monde réel. Les sujets des *news* sont variés, mais ce n'est pas pertinent pour nous. De manière plus générale, une nouvelle est une histoire relayée par un média. Les médias de communication des *news* sont pluriels. Pour les citer de manière non exhaustive, nous avons : la presse papier, la télévision, les radio-fréquences et les réseaux sociaux (Facebook, Twitter, etc.).

L'erreur est une opinion, un jugement ou une information non conforme avec la réalité ou la vérité telle que nous la percevons. Imaginons qu'un individu pense et déclare que tous les cygnes sont blancs et que toutes les personnes racisées sont des voleurs. Cet individu se trompe car il existe des cygnes noirs et des personnes racisées qui n'ont jamais volé de leur vie. L'erreur est inconsciente. Elle n'est pas intentionnelle. Lorsqu'elle est démasquée, elle tend à être corrigée. Imaginons qu'un journal, par inattention, présente une photo qui ne correspond pas au sujet de sa chronique, le journal publiera par la suite un *« Erratum »* [24]. Une nouvelle erronée est donc une histoire médiatisée qui n'est pas vraie. Elle ne correspond pas à la vérité. La nouvelle erronée est due à une erreur scientifique ou bien une erreur journalistique, c'est-à-dire une mauvaise manipulation de l'information par l'un de ces deux corps. Les médias sérieux corrigent leurs nouvelles erronées par des articles de démenti. Ceci fait partie du code de déontologie des journalistes pour l'honnêteté intellectuelle. Une *Fake News* [1] n'est pas une nouvelle erronée. Cependant les deux sont souvent confondues.

Parfois des nouvelles sont volontairement erronées. On dit alors qu'elles sont fausses. Il faut distinguer deux types de fausses nouvelles. Le point important ici est l'intention cachée.

Les fausses nouvelles dans un but bienveillant sont des satires ou des informations parodiques. Ces parodies imitent les médias. Elles volent même jusqu'à leur nom parfois (*Le Gorafi* est l'anagramme du *Figaro*). Mais au lieu de diffuser de vraies informations, les parodies proposent un contenu décalé, sarcastique ou qui relève du canular. Le but premier de ce genre de nouvelles est le divertissement. « Héritage de Stephen Hawking : Sa famille se dispute pour savoir qui obtiendra le trou noir dans son garage » [34]. « Barilla et Doliprane lancent des pâtes spéciales « fin de soirée » avec des vrais

1. En français, nouvelle truquée mais le terme *Fake News* est devenu tellement commun en français qu'il sera utilisé tel que, dans ce mémoire.

morceaux d'aspirine » [23]. Même si l'on trompe le lecteur, on ne cherche pas à lui nuire mais plutôt à le faire rire. Enfin, il y a toujours des exceptions où l'information de ces sources parodiques semble tellement crédible qu'elle est ensuite utilisée comme source fiable. « Christine Boutin cite Le Gorafi sur BFMTV [...] » [20] Les parodies donnent des informations délibérément fausses. La tromperie est totalement assumée et même souvent revendiquée.

Les fausses nouvelles qui ont pour but de volontairement tromper sont ce que l'on appelle des *Fake News*. Elles se distinguent de l'information erronée car elles ne sont pas le produit du hasard ou d'une mauvaise manipulation. Et elles se distinguent de la satire et de la parodie car elles ne sont ni assumées, ni revendiquées comme fausses. La *Fake News* provient d'une personne s'exprimant publiquement ou d'un ensemble de médias. Elle participe à la désinformation via les médias et les réseaux sociaux. Souvent, les *Fake News* sont écrites par des anonymes difficilement contestables et condamnables.

Pour prendre un site parmi tant d'autres, Secretnews est un site colportant de fausses informations, par exemple : « Alain Finkielkraut : « À ma mort, je léguerai tout mon patrimoine aux migrants. » » .[25] Il est dit dans cet article qu'Alain Finkielkraut aurait, dans une interview au *Figaro*, exprimé son souhait de léguer son patrimoine immobilier à la France pour y loger des migrants. Mais il n'existe aucune trace du-dit article. Bien sûr, cette information n'a pas d'auteur autre que le site qui ne donne aucun contact, et qui revendique que « toutes nos[leurs] sources sont vérifiées par un huissier assermenté » , mais aucun rapport ou mandat de justice n'apparaît sur le site. Les informations sur cette page paraissent tellement folles que l'on pourrait se demander s'il ne s'agit pas d'une satire.

Pendant l'« Émission politique » diffusée le 3 novembre sur la chaîne publique France 2, Jean-Luc Mélenchon fait plusieurs affirmations qui sont controversées. Il dit premièrement que la France détient le record du nombre de millionnaires. Or, cette information est fausse d'après l'étude de Capgemini, le « World Wealth Report 2017 » . Ensuite, il nie que l'opposition vénézuelienne se serait fait frapper devant leur Assemblée nationale, le 5 juillet 2017, alors que des médias vénézueliens comme *Noticias RCN* ont partagé des vidéos[31] de la répression qui a eu lieu. Il défend dans l'interview d'autres arguments en faveur du gouvernement de Nicolas Maduro. Mais la plupart sont factuellement fausses. Ici, Jean-Luc Mélenchon répand des *Fake News*.

Le 29 mars 2018, le site Direct Actu partage une photo qui fait polémique sur plusieurs médias peu scrupuleux. Il titre sur cette photo « Un politicien brésilien corrompu accroché à un poteau ! » [10]. Une citation de l'article dit : « Marcionólio da Costa Mendes a été attaché à un poteau par la population de Mata Grande, dans l'État de Alagoas. Accusé d'avoir volé dans les caisses publiques, le parlementaire a été innocenté par le tribunal de justice, ce qui a provoqué la colère des habitants de la commune. » Le fond de l'article reste à vérifier mais en tout cas, la photo partagée n'est pas celle de Marcionólio da Costa Mendes. Premièrement, le visage du personnage est brouillé donc pas de reconnaissance. Deuxièmement, la photo est celle d'une publicité pour une banque australienne : The financial brand.

Ici, Direct Actu partage un fausse information : la photo de l'article ne correspond pas à la réalité qui y est décrite.

(a) Un politicien brésilien corrompu accroché à un poteau !

(b) Doug Gets Tied to a Pole

FIGURE 1.1 Une photo servant à une campagne publicitaire et à une *Fake News*

1.1.2 Limites de la définition

Certes, une erreur peut arriver dans le traitement ou la création de l'information. Mais ces erreurs sont-elles légitimes ? La science progresse en faisant des erreurs. D'ailleurs, dans la définition de l'étude scientifique, l'erreur joue un rôle important. L'étude scientifique est la recherche perpétuelle de l'erreur. À défaut de nous apprendre ce qu'est la vérité, la science nous montre ce qu' « elle » n'est pas. Donc si l'erreur scientifique est légitime au bon fonctionnement de la science, peut-on en dire autant de l'erreur journalistique ? Le journaliste doit faire un compte rendu exhaustif, objectif et vraisemblable du sujet qu'il traite. Si une erreur factuelle ou non scientifique vient se faufiler dans son article, c'est qu'il a mal fait son travail.

Par exemple, on voit toujours des médias relayer l'information que les vaccins causent l'autisme chez l'enfant. Cette croyance persiste après une publication d'Andrew Wakefield [2]. Ses publications furent démenties à de multiples reprises par des autorités compétentes, mais les médias de grande audience ont continué à transmettre ce message de méfiance vis-à-vis des vaccins. Le journaliste ne respecte pas alors sa déontologie. Les médias traditionnels continuent de relayer le discours erroné des antivax. [3] , qui fait couler beaucoup d'encre. Mais cette controverse est de la désinformation pure et d'une grande malhonnêteté intellectuelle. On retrouve sur le site de fausses informations StopMensonge, cet article : « Le Président Duterte banni[sic] les vaccins aux Philippines : "Les Vaccinations provoquent l'Autisme" » [12]. L'information est peut-être fausse mais s'il s'avérait que le Président ait réellement pris une telle décision, il aurait été victime d'une des *Fake news* les plus répandues. Mais cette nouvelle est donnée pour introduire un doute dans l'esprit du lecteur sur la vaccination. Et quand, par la suite, les médias de forte audience publient des articles comme celui de *FranceInfo* : « Vaccination et autisme : Nous réclamons justice et réparation pour nos enfants blessés

2. Andrew Wakefield est un ancien chirurgien et chercheur médical britannique connu pour ses affirmations frauduleuses sur le vaccin ROR et l'autisme. Il fut radié de l'ordre des médecins britanniques en mai 2010 pour défaut à son devoir de consultant responsable.

3. Partisans de la controverse sur la vaccination qui remettent en cause son efficacité et la sécurité de certains vaccins.

par leur vaccin » [32], les indécis qui avaient lu des articles comme celui de StopMensonge croient en cette désinformation persistante malgré les méta-analyses qui infirment ce lien de causalité. Les médias sont toujours les vecteurs de transmission de cette *Fake News*.

1.2 Pourquoi produire des *Fake News* ?

1.2.1 Le contexte des *Fake News*

L'invention qui a le plus contribué à l'essor des médias est certainement l'imprimerie. Elle nous fait passer des histoires orales à la presse. L'information est figée sur du papier. Elle n'est plus perdue ou transformée par le locuteur de l'histoire. À partir de ces documents, on peut faire des versions officielles et approuvées par une autorité. L'information fut pendant très longtemps partagée de manière verticale. La source d'autorité de la connaissance était plus ou moins légitime et compétente. La connaissance était donnée par les médias et leur vision. Il n'y avait pas de sources alternatives ou contestataires.

Puis Internet fut créé ! Internet permet un partage des connaissances horizontal. Toutes les personnes disposant d'une connexion au réseau peuvent contribuer à la connaissance générale. Chacun peut écrire, partager et relayer des informations. Les médias se sont développés sur Internet et surtout, ils s'y sont multipliés. La pluralité du web permet de mieux recroiser ses sources. Nous ne sommes plus dans la vision dogmatique des grands groupes qui possèdent les chaînes télévisuelles ou les organes de presse.

Nous avons acquis une énorme liberté d'expression avec l'avènement d'Internet comme média souverain de la pluralité. Mais ce nouveau régime pluriel a aussi des désavantages. La démocratie de la connaissance permet aux profanes de s'exprimer sur des sujets de pointe. Ainsi, nombreux sont les opinions et les préjugés qui sont travestis en pseudo-vérités. Paradoxalement, la liberté d'expression et la facilité d'accès à Internet participent grandement à la désinformation globale.

De plus, la multiplicité de perspectives qu'engendre Internet pose un problème à notre compréhension du monde. En effet, comme le dit Michael Lynch, utiliser Internet c'est « connaître plus et comprendre moins [...] » [4]. Lynch conteste la notion largement acceptée qu'Internet est un avantage parce qu'il rend plus d'informations disponibles à plus de personnes, plus rapidement et facilement.

Ainsi, les *Fake News* circulent librement sur la plate-forme qu'est Internet. Elles dégradent des informations primaires qui peuvent être complexes pour faire du sensationnalisme. Les réseaux sociaux sont leurs meilleurs médias de transmission. Le phénomène est assez nouveau. En effet, si nous observons les tendances de recherches sur Google, le terme « *Fake News* » a connu sont apogée en 2017 et depuis il est couramment utilisé.

4. Citation : *Knowing more and understanding less in the age of big data* sous-titre du livre *The internet of us*, de Michael Lynch ISBN-10 : 0871406616

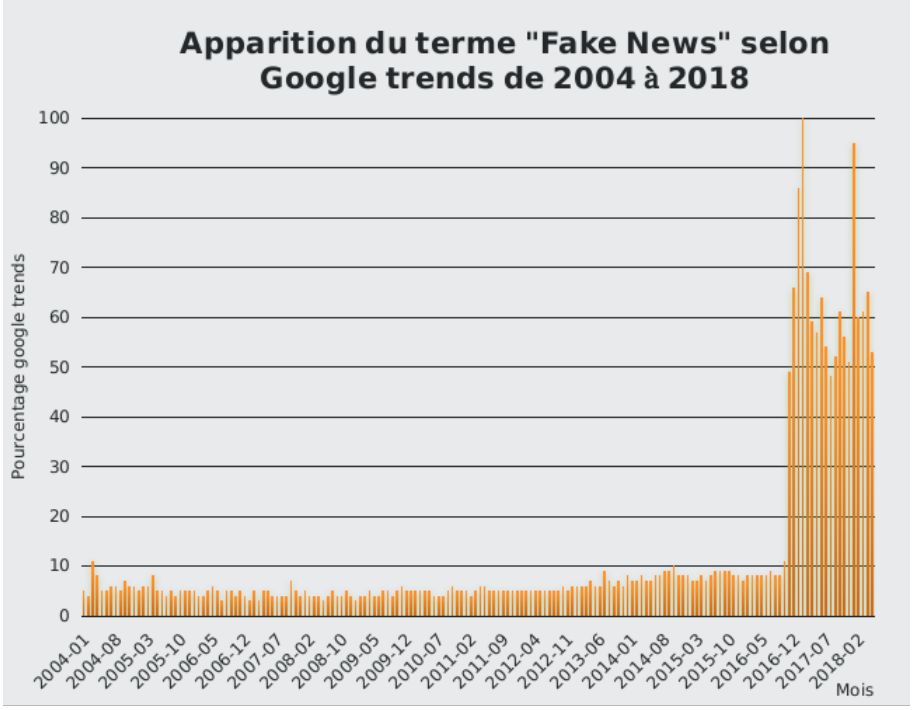


FIGURE 1.2 Évolution de l'intérêt pour la recherche « *Fake News* »
5

1.2.2 Le système monétaire des *Fake News*

Le *clickbait* désigne un contenu web qui vise à attirer le maximum de passages d'internautes afin de générer des revenus publicitaires en ligne. Le clickbait affiche des gros titres racoleurs. Il est souvent mensonger et sensationnel. L'exactitude et les sources de l'article sont inexistantes. Le but du *clikbait* est d'être partagé massivement sur les réseaux sociaux.

Maintenant que la presse virtuelle est multiple, pour être rentable, elle doit générer une offre publicitaire non négligeable. Le *clickbait* est un effet pervers d'Internet. Les possesseurs de contenu peuvent gagner de l'argent par le biais de la publicité. Cela crée des nouveaux systèmes monétaires.

La ville de Vélès en Macédoine est devenue le temps des élections américaines une « [...] usine à *Fake News* [...] » [28]. En tous cas, c'est ce que titre Sputniknews, et sur ses pages il propose l'interview de Dimitr, étudiant en Web-design qui cherchait à se faire de l'argent facilement. « À un moment donné [Dimitr] et les [autres] étudiants ont compris qu'il était possible de gagner de l'argent avec la publicité en écrivant des nouvelles selon une équation simple : plus il y a de lecteurs, plus cela rapporte de l'argent.[...] La plupart d'entre eux publiaient des articles pour soutenir Trump, ou sur lui. »

Les *Fake News* et le *clickbait* associés répondent à la crise profonde de la presse papier au profit des réseaux sociaux comme médias. De plus, les *Fake News* alimentent une méfiance envers les médias traditionnels. Elles font croire avec leurs « faits alternatifs » qu'on tente de cacher quelque chose à la

5. L'évolution de l'intérêt pour une recherche est ainsi définie par Google : « Les résultats reflètent la proportion de recherches portant sur un mot clé donné dans une région et pour une période spécifiques, par rapport à la région où le taux d'utilisation de ce mot clé est le plus élevé (valeur de 100). Ainsi, une valeur de 50 signifie que le mot clé a été utilisé moitié moins souvent dans la région concernée, et une valeur de 0 signifie que les données pour ce mot clé sont insuffisantes. »

population. Ainsi, ce nouveau type de média essaie de conserver au plus haut niveau son lectorat pour conserver des rentrées d'argent constantes.

1.2.3 Raisons idéologiques

Les raisons idéologiques de la production des *Fake News* ne manquent pas, en particulier en politique, où elles sont utilisées à tout va. Un exemple probant est le Pizzagate. En résumé le Pizzagate est une « théorie » conspirationniste prétendant qu'il existe un réseau de pédophilie autour de John Podesta, l'ancien directeur de campagne d'Hillary Clinton. Cette histoire fut rapidement démentie par les services de police et une majorité des médias américains. Mais les conséquences ne furent pas négligeables. Un fusillade, heureusement sans blessés, a eu lieu dans la pizzeria où étaient soi-disant séquestrés les enfants du réseau pédophile de Podesta.

1.3 Les origines des *Fake News*

1.3.1 Des médias négligents

La négligence et l'unicité des médias font que les *Fake News* peuvent se propager facilement. Comme nous l'avons dit précédemment, les médias traditionnels traversent une crise. La production de contenu doit être faite le plus rapidement possible. Dans le but de répondre à la course à l'information de plus en plus grande et déloyale, la qualité des articles est revue à la baisse. Certains journalistes ne font pas le travail de vérifier leurs sources afin d'accélérer le procédé de publication.

1.3.2 Des organisations malveillantes internationales

L'appât du gain facile que sont les *Fake News* a suscité des vocations. Comme cité précédemment, une partie de la jeunesse de Vélès s'est ainsi spécialisée dans la création de *Fake News*, attirée par de juteux revenus publicitaires. La ville s'est transformée en fabrique à *Fake News* pendant l'élection américaine[14]. Mais des sites comme InfoWars voient aussi naître des *Fake News*.


File: tiphaine macron .jpg (424 KB, 833x538)

MEME Macron out of existence - Tiphaine Edition

Anonymous (ID: D1g4l+nT) 04/24/17(Mon)04:17:08 No.122527652 [Reply]

We need to stop this creep Macron from cucking up Europe, Le Pen is France's, and by extension, Europe's, last hope. We've got two weeks to wipe that Rothschild-smile off his face.

The left loves him because he wifed up his old teacher, Brigitte, who is like 30 years older than him, it plays into their "muh stronk, older woman can get hot young toy boys"-fantasy. Brigitte has got a 30 year old daughter from a previous marriage Tiphaine Auziere. Wouldn't it be funny if it got out that Macron had engaged, or tried to engage in improper conduct with the young (arguably) hot daughter of his strong semi-senile wife in the past - that his wife is a woman scorned, but hush money from (((the establishment))) was doused on the matter to keep it a secret. (that last part about the hush money might be a stretch, but hey)

There have already been allegations about Macron being a closeted homosexual so the idea that he is a sexual pervert is already planted in the peoples' heads...people might buy it and that would taint him into oblivion.

Well what do I know anons, I'm just a Swedecuck grasping for straws here but something must be done in these coming two weeks and this is the best angle I came up with for now, the frog-anons would obviously have to help spread this in frog.

FIGURE 1.3 Capture d'écran d'un fil de discussion sur 4chan pour lancer une rumeur de pédophilie sur Macron.

Des organisations qui travaillent dans l'ombre des plus grands forums sont à la base de la conception des *Fake News*. En effet, l'étude de Savvas Zannettou et al, 2017[38] montre comment les forums leaders mondiaux que sont 4chan et reddit sont en partie responsables de la création de rumeurs.

1.4 La propagation des *Fake News*

1.4.1 Les réseaux sociaux : médiations des *Fake News*

Selon Maksym Gabielkov et al, 2016[13], 59% des liens partagés sur Twitter n'ont jamais été cliqués. En d'autres termes, la plupart des gens semble retweeter des nouvelles sans jamais les lire. Les personnes qui partagent sans lire des articles propagent certainement des *Fake News* sans le savoir.

Pour évaluer une *Fake News*, nous n'avons pas besoin de diplôme. En effet, les déterminismes sociaux ne sont pas suffisants pour prévoir le partage de *Fake News*. L'éducation, le sexe, l'âge, etc. ne sont pas des critères distinctifs. Nous pouvons inférer cela de l'étude « Le conspirationnisme dans l'opinion publique française » [21]. Dans cette étude, on affirme qu'un français sur quatre sans distinction de sexe ou d'âge croyait à au moins une théorie du complot. L'échantillon de 1 250 personnes est représentatif de la population française. Si intrinsèquement les personnes croient à des théories qui ne sont pas la vérité, il est peu étonnant que les *Fake News* soient autant partagées sur les réseaux sociaux.

Nous pourrions nous demander « Pourquoi tout le monde - ou presque - partage des *Fake News* ? » [30] La réponse se trouve en partie dans le « Digitial News Report 2017 » où il est dit que « 51% des adultes Américains s'informent sur les réseaux sociaux, contre 38% en France » .

Nous voyons alors que personne n'est à l'abri des *Fake News*. Seuls la pensée critique et le recul nous permettraient d'être protégés des *Fake News*.

1.4.2 Le biais de confirmation

Les biais de confirmation sont un aspect déroutant de la pensée humaine. Nous pourrions penser que l'homme a acquis une pensée analytique développée pour arriver à notre niveau d'intelligence. Et pourtant, il est soumis au biais de confirmation. Ce biais cognitif consiste à privilégier les informations confirmant nos idées préconçues ou nos hypothèses. De plus, il nous fait aussi négliger les informations jouant en défaveur de nos conceptions.

Ainsi, les personnes tenantes de la thèse que « des extraterrestres gouvernent le pays » sont plus enclines à croire la thèse « des reptiliens au pouvoir » plutôt que la thèse selon laquelle « Barack Obama est un être humain » [21].

Les *Fake News* relayant souvent des informations conspirationnistes, il est facile pour un adepte de croire celles-là plutôt que de croire des versions officielles. Le biais de confirmation agit pour tous les sujets confondus. Il n'est pas uniquement cantonné au conspirationnisme. Les *Fake News* utilisent ce biais de manière idéologique pour renforcer nos croyances et rendre leur information plausible.

« Des anti-immigration prennent des sièges de bus pour des femmes voilées » [8]. Sinder Beyer, un politicien norvégien, dénonce l'utilisation abusive d'une photo de siège de bus et l'idéologie raciste que prônent les anti-immigration sur un de leurs postes Facebook. Nous avons bien ici un renforcement des croyances prenant ses bases dans une *Fake News*. C'est parce que ce groupe anti-immigration pense que la Norvège est envahie petit à petit par des musulmans qu'il confond des sièges d'autobus avec des femmes en burqa.

FIGURE 1.4 Post de Sindre Beyer pour dénoncer l'erreur des anti-immigration.

1.4.3 Tri de l'information selon Kahneman

La thèse centrale de Daniel Kahneman, dans son livre *Système 1 / Système 2 : Les deux vitesses,* est qu'il y a une dichotomie entre deux modes de pensée. Le système 1 est rapide, instinctif et émotionnel, alors que le système 2 est plus lent, plus réfléchi et plus logique. Il définit les biais cognitifs associés à chacun de ces modes de pensée. Il montre que l'on donne une trop grande importance au jugement humain.

Selon Kahneman nous nous reposons plus souvent sur le système 1, ce qui expliquerait notre partage de *Fake News* quand nous sommes émotionnellement impliqués.

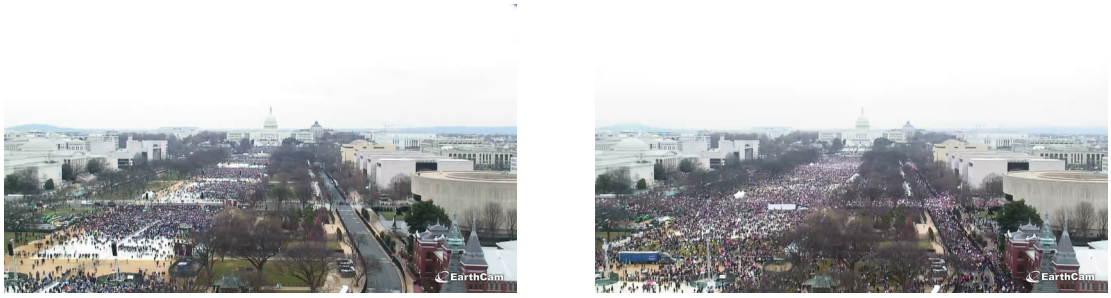
Imaginons qu'une personne antivax « ouverte d'esprit » voie cet article apparaître sur son fil d'actualité Facebook, et qu'on lui demande d'évaluer la fiabilité de l'information suivante : *« NOW IT'S OFFICIAL: FDA Announced That Vaccines Are Causing Autism! »* [2]. Cette personne devant traiter l'information rapidement va utiliser un argument d'autorité de sa communauté pour nous répondre que cette information est vraie. Elle a utilisé un raisonnement intuitif (système 1). Maintenant on lui demande de croiser les informations, remonter les sources et d'évaluer sérieusement les preuves avec des calculs comparés avec des expériences témoins. Si cette personne fait bien sont travail, elle devrait tomber sur un article du genre *« No, the FDA didn't hide information linking vaccine to autism »* [35], et trouver peu de différences, statistiquement non-significatives, dans le pourcentage d'autistes chez les enfants vaccinés par rapport aux enfants non-vaccinés. Ici, elle aura eu un raisonnement analytique et scientifique, mais qui lui aura pris beaucoup de temps (système 2). Et si cette personne est vraiment ouverte d'esprit et de bonne foi, elle devrait changer un peu ses opinions sur les vaccins, car elle aura eu les preuves tangibles sous les yeux. Cela ne veut pas dire qu'elle doit totalement changer

d'avis mais qu'elle devra considérer que les vaccins sont moins nocifs ou du moins qu'ils ne causent pas l'autisme.

1.5 Les risques des *Fake News*

1.5.1 Les risques politiques

L'institutionnalisation des *Fake News* est l'un des plus gros risques politiques. En effet, rien de pire qu'une *Fake News* qui a pour pseudo-autorité un Etat. Un Etat niant les faits devient répressif.



(a) L'inauguration de Trump (b) Women's March

FIGURE 1.5 Deux photos d'événements du National Mall à Washington

L'inauguration du Président Trump et la Women's March aux Etats-Unis avaient fait beaucoup de bruit. Les partisans pro-Trump étaient soi-disant en plus grand nombre que les partisans de la Women's March. Nous pourrions croire la version officielle. Mais apparemment, sans comptage, il y avait quand même plus de participants à la Women's March. Kellyanne Conway, Conseillère en Communication à la Maison Blanche avait tenté de faire passer cela pour des faits alternatifs, ce qui n'a pas beaucoup de sens. Certes, nous appréhendons tous le réel de manière différente, mais ce n'est pas une raison pour faire du relativisme et conclure à des faits alternatifs. De plus, comment définir des faits alternatifs ? Prendre des faits et les dénaturer à des fins idéologiques ne nous donne pas raison sur la réalité. En somme, les *Fake News* pourraient servir à cacher des scandales politiques.

D'autres *Fake News* institutionnalisées peuvent servir à des particuliers pour qu'ils puissent s'enrichir, ou à ne pas être mis à contribution pour résoudre un problème. Par exemple, nier le réchauffement climatique est très pratique quand on est l'un des pays les plus polluants au monde.

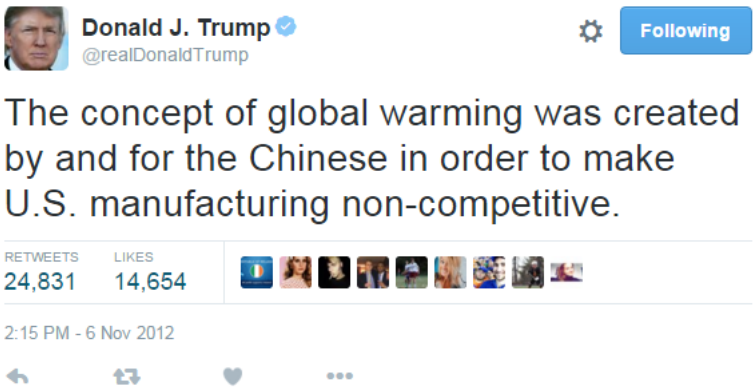


FIGURE 1.6 Tweet de Donald Trump niant le réchauffement climatique.[17]

« Donald Trump sort les États-Unis de l'Accord de Paris sur le climat » [26]. Comme il avait promis dans sa campagne, Donald Trump ne souhaite pas toucher aux emplois américains. Il ne va pas prendre des mesures écologiques en ce qui concerne l'industrie des États-Unis. Ainsi, il décide de ne pas se joindre aux efforts des différentes nations pour diminuer la température globale de deux degrés Celsius.

1.5.2 Les risques sanitaires

Beaucoup de *Fake News* portent sur les domaines sanitaires. Comme nous l'avons vu précédemment avec les vaccins, les *Fake News* développent une méfiance pour la médecine conventionnelle. Pour reprendre l'exemple des vaccins, il faut savoir que des cas de rougeole mortels sont réapparus ces dernières années. Par exemple : « Les cas de rougeole ont doublé en une année » [27] ou bien encore « La rougeole continue à être mortelle en France » [7]. Ne pas se vacciner entraîne un baisse de la couverture vaccinale. Le rapport bénéfice/risque est plus que positif pour la vaccination. Aucun des effets indésirables notoires suspectés n'a trouvé un protocole testable positif. Nous n'avons que des affirmations pseudo-scientifiques et des hypothèses contre. En somme, la désinformation, autour de la médecine notamment, peut coûter la vie.

1.5.3 Détournement des news vers des sujets clivants

Comme nous avons pu le voir précédemment les *Fake News* peuvent être colportées par des personnes convaincues de leur véracité. Le but ultime d'une *Fake News* est d'être considérée comme une vraie information. Les *Fake News* détournent l'attention de la population vers des problèmes factices et souvent résolus depuis des années. En somme les *Fake News* sont une perte de temps. Elles nous empêchent de nous focaliser sur de vrais problèmes, car elles nous obligent à démystifier des faits irréels.

1.6 Légitimité du projet et motivations

Comme nous avons pu le voir dans les sous-sections précédentes. Les *Fake News* sont de fausses nouvelles qui cherchent à tromper volontairement. Elles participent pour la majorité à une pollution du web, encouragée par le système financier de la publicité en ligne. Elles servent parfois un propos idéologique avec de mauvais arguments. Cela les rend moins crédibles pour ceux qui découvrent le pot aux roses. Elles sont produites par des militants anonymes anti-intellectuels prônant la désinformation sur de sombres forums. Elles inondent les réseaux sociaux d'inepties et sont partagées en masse sans être lues. Elles véhiculent des titres avec de fausses idées alors qu'un simple coup d'œil au corps de l'article rend le propos infondé. De plus, elles coûtent la vie à certains. Et pour finir, elles peuvent devenir un instrument politique redoutable et anti-démocratique.

Pour toutes ces raisons, il est légitime de vouloir combattre le phénomène des *Fake News*.

1.7 Comment détecter une *Fake News* ?

1.7.1 Comment vérifier des informations ?

Il y a deux moyens de détecter une *Fake News*. Par soi-même en cherchant les indices du canular, ou en utilisant un moteur de recherche de canular sur un domaine Internet.

Premièrement, voyons de manière non exhaustive, quelques techniques pour détecter une *Fake News* :

Avant de partager, il faut se questionner sur ce qui est raconté dans l'article et vérifier les sources. On est responsable de ce que l'on partage.

Est-ce une information ? Il faut se poser différentes questions : est-ce que cela a un intérêt public ? Est-ce factuel ? Est-ce vérifié ? Cela permet de distinguer les avis et les rumeurs des informations.

Ce site est-il fiable ? A-t-il un onglet « À propos » sur sa page ? Est-il parodique ? Quelles sont les sources de ce site ?

Beaucoup de techniques spécifiques pour chaque média existent ! Nous ne pouvons pas être exhaustif ici.

Les *Fake News* ont fait apparaître de nouveaux sites spécialisés dans la détection de canulars. Par exemple, en français, il existe le Décodex proposé par le journal *Le Monde*. Ce site répertorie les autres sites selon leur fiabilité. En anglais, il existe le site Internet Polifacts de vérification des faits, qui vérifie la véracité des promesses et engagements pris par les politiques américains

1.7.2 Méthode informatique actuelle : *Stance Detection*

Cette tâche peut aussi être résolue avec un succès modeste par l'apprentissage automatique. En effet, une des réponses au phénomène des *Fake News* est l'apparition de réseaux neuronaux complexes qui permettent de manière partielle de repérer les *Fake News*. Ce repérage passe par une détection des partis pris (*Stance Detection*).

Ainsi, dans la deuxième partie de ce mémoire, nous allons voir plus en détails ce qu'est la *Stance Detection*. Au travers de deux tâches partagées, nous allons découvrir les techniques pour l'utiliser.

Dans la troisième partie, nous nous essaierons à la *Stance Detection* et nous donnerons une analyse comparative de nos résultats par rapport à l'état de l'art.

Chapitre 2

État de l'art : *Stance Detection*

Dans ce chapitre, nous allons présenter la *Stance Detection* et lui donner une définition rigoureuse. Nous discuterons les différentes tâches partagées qui ont eu lieu autour de la *Stance Detection*. Notamment, nous présenterons les corpus et les méthodes de classification choisies.

2.1 Vers une définition

2.1.1 Formalisation de la *Stance Detection*

La *Stance Detection* ou en français la détection de parti pris [1] est la méthode qui permet de déterminer si un énoncé **E** par rapport à une cible **C** donnée est en accord ou en désaccord avec la cible. On l'utilise aussi parfois pour déterminer si l'énoncé discute sans parti pris de la cible. C'est-à-dire que **E** parle de **C** mais on ne trouve aucun indice soit en faveur, soit en défaveur la cible. Par extension, si on arrive à déterminer la cible, on peut connaître les énoncés qui n'y correspondent pas. Ce type d'énoncé indépendant ne donne aucun indice de prise de parti et surtout aucun indice de la cible en général. La cible et l'énoncé ne partagent aucun lien direct ou indirect. Appellons **R** la relation entre **E** et **C**. Ainsi, la détection de parti pris nous permet de repérer ces quatre cas de figure :

E est pour C quand l'énoncé montre un ou des indices en faveur de la cible.

E est contre C quand l'énoncé montre un ou des indices en contradiction avec la cible.

E discute C quand l'énoncé donne une ou des informations sur la cible sans donner d'indices de parti pris comme dans les deux cas pécédents.

E est non-lié à C quand l'énoncé ne donne aucune information par rapport à la cible en général.

Les deux derniers cas ne détectent aucun parti pris mais ils restent pertinents pour la détection de parti pris qui doit intégrer une limitation à un sujet donné. Une telle détection implémente donc un module de détection des relations.

Imaginons par exemple une cible **C** composée de cet ensemble de phrases :

C1 Adam vend des pommes.

1. Vous remarquerez que, pour la cohérence et pour la consistance de ce document, nous n'emploierons plus que la formulation « détection de parti pris » au lieu de « *Stance Detection* » sauf pour les titres.

C2 Charles rencontre Adam.

C3 Charles achète une pomme à Adam.

Nous pouvons ainsi formuler des énoncés **E** qui exemplifient chacune des relations **R** possibles :

R : est pour Charles a acheté une pomme.

R : est contre Charles n'a pas acheté une pomme.

R : discute Charles et Adam aiment les pommes.

R : est non-lié Jean est à la plage.

2.1.2 Domaine de la *Stance Detection*

Nous avons beaucoup parlé d'indices dans la partie précédente. Questionnons leur nature. La détection de parti pris se base sur la relation entre la cible et l'énoncé. La nature de cette relation est sémantique. Les traits qui permettent d'unir l'énoncé et la cible pour déterminer le parti pris doivent alors aussi entretenir une relation sémantique. Si nous reprenons l'exemple de nos phrases cibles **C** au-dessus, les énoncés **E1** et **E2** suivants devraient avoir une relation **R : est pour** :

E1 Charles a dépensé de l'argent pour une pomme.

E2 Charles a acheté un fruit.

La détection de parti pris doit prendre en compte la synonymie comme dans **E1**. « Dépenser de l'argent pour » est synonyme périphrasé d' « acheter » . La détection de parti pris doit aussi gérer les liens d'hyponymie ou d'hyperonymie comme dans **E2**. « fruit » est un hyperonyme de « pomme » , et inversement « pomme » est un hyponyme de « fruit » . On verra plus tard que certaines relations syntaxiques peuvent être utiles de manière localisées dans certaines tâches.

Il apparaît un problème dû à la sémantique : il faut s'accorder au préalable sur le sens des mots et sur ce que désignent nos cibles et leurs rapports avec l'énoncé. En effet, la nature polysémique des mots peut parfois porter à confusion. Si nous prenons l'énoncé **E3** suivant :

E3 Charles a une grosse pomme d'Adam.

Ici, on ne sais pas si l'on parle d'une grosse pomme qu'Adam a donné à Charles ou bien si on parle de la proéminence laryngée particulièrement grosse chez Charles. Ainsi, dans d'autres énoncés **E**, pour clarifier le contexte d'énonciation, nous ne pouvons pas décider entre une relation **R : est non-lié** ou **R : est pour** (voir **R : discute** vu qu'il n'y a pas d'indice à propos d'achat). Pour constituer un corpus de prise de position par rapport à un énoncé, il faudra trouver la manière la plus objective de qualifier les termes afin de faire de meilleures relations entre l'énoncé et la cible.

2.1.3 La genèse de la *Stance Detection*

Une des premières publications sur la déctection de prise de parti était « *Cats Rule and Dogs Drool ![...]* » (voir Pranav Anand et al, 2011[3]) qui cherchait à classer la prise de position dans des débats en ligne sur des sujets variés. Leur but était de montrer que les débats idéologiques comportent une plus grande part de messages de réfutation par rapport aux autres discutions thématiques (sur d'autres fils de forums).

La publication montrait aussi qu'il est beaucoup plus difficile de classer ces posts de réfutation, aussi bien pour les humains que pour les classificateurs automatiques formés à la détection de prise de parti.

Les chercheurs ont créé leur corpus à partir de 1113 débats bi-partiaux (soit 4873 posts) dans 14 sujets différents du site ConvinceMe.net. Pour annoter le corpus, les chercheurs ont demandé à neuf participants de mettre une étiquette sur différentes parties de débat sur le site. Il est intéressant de noter que les annotateurs n'ont eu que 0.73 d'exactitude croisée [2] dans la classification des réfutations (tous sujets confondus). Ainsi, le problème sémantique est réel. Les annotateurs ne tombent pas d'accord sur la relation sémantique de certains énoncés par rapport à leurs cibles. Le but d'un système automatique de classification des réfutations sera donc de s'approcher au plus de ce pourcentage d'exactitude pour représenter au mieux le classement humain général.

L'équipe d'Anand a fait plusieurs modèles de système utilisant des traits différents (n-grams, la ponctuation répétée, LIWC [3], dépendance syntaxique...). Ces modèles utilisaient soit une implémentation de Naive Bayes, soit une implémentation de JRip [4] en fonction des traits choisis.

Les résultats des modèles varient en fonction des sujets entre 0.59 et 0.69 d'exactitude pour la détection de réfutation. De plus, aucun modèle ne se départage des autres. Tous ont plus ou moins réussi selon un sujet différent. La moyenne est de 0.63 pour la détection de réfutation. C'est dix points de moins que l'exactitude humaine (qui varie entre 0.66 et 0.94).

Premièrement, cette publication nous montre combien il est ardu de se confronter à la sémantique. En théorie, le sens d'un énoncé devrait être univoque et susciter une seule relation universelle entre la cible et l'énoncé. En fait, nous - les annotateurs humains - sommes tous soumis à des biais, des opinions, des préjugés qui ne permettent qu'un recouvrement subjectif de la relation entre cible et énoncé.

Deuxièmement, cet article montre la difficulté de la tâche. Aucun modèle possédant des traits particuliers n'a été supérieur dans tous les sujets confondus.

Nous avons présenté un des premiers travaux sur la détection de parti pris. Par la suite, nous nous limiterons à des travaux plus récents. Ceux-ci s'intéressent à notre sujet : les *Fake News*.

2.1.4 Application générale et relative au *Fake News*

Nous allons voir dans les sections suivantes les différentes utilisations de la détection de parti pris à travers deux tâches partagées. Premièrement, dans la section n° 2, nous verrons l'utilisation de la détection de parti pris à l'intérieur de *tweets* sur des sujets polémiques ; grâce à la ***SemEval-2016 Task 6***. Puis, deuxièmement, dans la section n° 3, nous verrons l'utilisation de la détection de parti pris pour la détection de *Fake News* en se basant sur au ***Fake News Challenge***. Nous proposerons dans la troisième partie de ce mémoire une contribution originale à la tâche que nous discutons dans la section n° 3.

2. L'exactitude croisée est une technique de validation de modèle permettant d'évaluer comment les résultats d'une analyse statistique seront généralisés à un ensemble de données indépendant.

3. *Linguistic Inquiry and Word Count dictionaries*, référence en matière d'analyse de texte informatisée.

4. JRip est un classificateur basé sur des règles qui produisent un modèle compact adapté à la conception rapide d'applications.

2.2 *SemEval-2016 Task 6*

2.2.1 Description générale de la tâche

Cette tâche porte le nom de « *Detecting Stance in Tweets* » (Saif M. Mohammad et al, 2016[19]). Le but de la tâche est de déterminer la relation **R** entre un tweet **E** et une cible **C**.

Ici, nous avons trois classes possibles pour déterminer le parti du tweeteur par rapport à son tweet :

favor s'il est en faveur de la cible.

against s'il est contre la cible

neither si il n'est ni en faveur de la cible, ni il n'est contre la cible.

Exemple direct :

C Hillary Clinton

E Hillary Clinton has some strengths and some weaknesses.

R neither

Exemple indirect :

C legalization of abortion

E A foetus has rights too ! Make your voice heard.

R against

La tâche se divise en deux sous-tâches :

Sous-tâche A

La sous-tache A est supervisée. Elle porte sur cinq différents sujets (*'Atheism'*, *'Climate Change is a Real Concern'*,*'Feminist Movement'*, *'Hillary Clinton'*,*'Legalization of Abortion'*). Elle contient 4163 couples **C/E** labélisés **R**. On réserve 30% pour l'ensemble de test et le reste pour l'ensemble d'entraînement.

Sous-tâche B

La sous-tache B est non-supervisée. Elle porte sur un seul thème (*'Donald Trump'*). L'ensemble test est constitué de 707 tweets **E**. Aucun ensemble d'entraînement n'a été donné. Mais un ensemble de 78 000 tweets non-labellisés à propos de Donald Trump était disponible.

Évaluation

Pour l'évaluation des systèmes on utilise le F1-score moyen ; calculé ainsi :

$$F_{avg} = \frac{F_{favor} + F_{against}}{2} \tag{2.1}$$

2.2.2 Les participants

Il y a eu 19 dépôts pour la sous-tâche A et 9 dépôts pour la sous-tâche B. Premièrement, nous parlerons de la réussite pour la tâche en général. Deuxièmement, nous discuterons seulement de quelques dépôts intéressants. Et troisièmement, nous comparerons les différents protocoles vus.

Dépôts en général

Parmi les 19 dépôts de la sous-tâche A, aucun système n'a dépassé les baselines. En effet, Saif M. Mohammad et al.[36] ont fourni quatre baselines pour la sous-tâche A. La première baseline naïve donnait le label majoritaire à toutes les entrées (*Majority class*). Les trois suivantes utilisent un ou plusieurs classificateurs linéaires (SVM) pour labelliser les entrées. La deuxième utilise 5 classificateurs SVM (un pour chaque sujet) sur les vecteurs d'unigrams de la combinaison de **C** et **E** (*SVM-unigrams*). De la même manière, la troisième a aussi 5 SVM mais utilise comme dimensions de vecteur : les 1-2-3-grams pour les mots et les 2-3-4-5-grams pour les caractères (*SVM-ngrams*). Pour finir, la quatrième baseline utilise un seul SVM sur les mêmes dimensions de vecteur que la troisième baseline (*SVM-ngrams-comb*).

En revanche, 7 des 9 équipes ont réussi à battre les baselines de la sous-tâche sur B. Ces baselines reprennent certaines baselines de la sous-tâche A, à savoir la première *Majority class* et la quatrième *SVM-ngrams-comb*.

Pour la sous-tâche A, la plupart des équipes ont utilisé des fonctions de classification de texte standard telles que n-grams et vecteurs de mots, des lexiques de sentiments. Certaines équipes ont interrogé Twitter sur des hashtags, pour marquer des prises de parti plus déterminantes. Certaines ont entraîné leurs systèmes à partir de vecteurs de Google Actualités ou directement des corpus Twitter.

Et pour la sous-tâche B, certaines équipes ont très bien détecté les tweets en faveur de Trump, grâce aux tweets qui se trouvaient dans le corpus de Clinton en inversant leur valeur.

Dépots particuliers

Ici, de manière succincte, nous allons décrire les différents modèles des meilleurs systèmes pour les différentes tâches.

MITRE, 1er pour la tâche A[39]

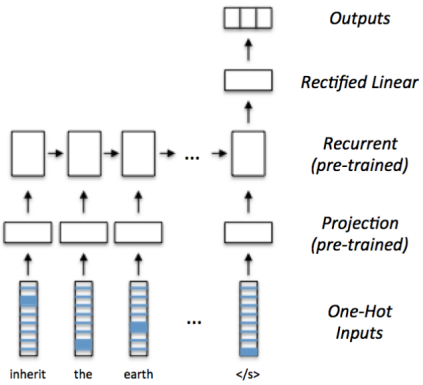


FIGURE 2.1 Le réseau neuronal récurrent de MITRE pour la détection de parti pris

L'approche de détection de prise de parti de MITRE utilise un réseau de neurones récurrent organisé en quatre couches de poids. Chaque mot est projeté comme une entrée dans une couche d'*embbeding* de 256 dimensions (créé avec word2vec), qui alimente en une couche récurrente contenant 128 LSTM. La sortie du terminal de cette couche récurrente est densément connectée à 128 classificateurs linéaires. Cette même couche est entièrement connectée à un softmax tri-dimentionnel dans laquelle chaque unité représente l'une des classes de sortie : FAVOR, AGAINST ou NONE.

pkudblab, 2ème pour la tâche A[37]

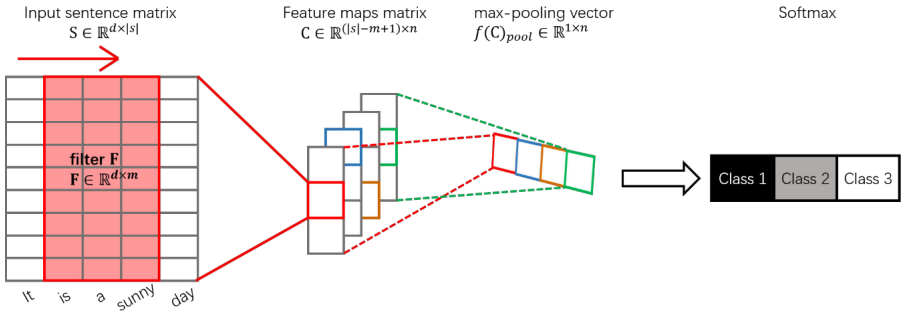


FIGURE 2.2 Architecture principale du réseau de neurones convolutionnels de pkudblab

L'architecture principale de pkudblab est un réseau de neurones convolutionnels. On peut en décrire 5 composants majeurs :

Une table de correspondance est une énorme matrice d'embbeding de mots. Chaque colonne de la table correspond à un mot. Chaque mot incorporé dans cette table a été pré-formé par des vecteurs de word2vec (Mikolov et al., 2013). Ces vecteurs sont formés sur une partie de l'ensemble de données Google News.

Une matrice d'entrée représente une phrase d'entrée et la longueur de la phrase.

La couche convolution a pour but d'extraire des patterns, de sorte que certaines formulations abstraites communes soient représentées.

Le Pooling layer a pour but de simplifier l'information dans le sortie de la couche convolutionnelle.

La Couche de sortie softmax est entièrement connectée et est destinée à la classification.

Pour la sous-tâche A, ils ont séparé les ensembles de données en cinq sous-ensembles. Ils ont donc entraîné de manière séparée les modèles sur les différents sujets.

pkudblab, 1er pour la tâche B[37] Pour la sous-tâche B, ils ont utilisé le même modèle que pour A. Pour entraîner leur modèle, ils établissent un ensemble de données de formation à deux classes (favor, against) à partir du corpus du domaine officiel en fonction de plusieurs expressions spéciales.

Discussions comparatives

Baseline	$\mathbf{F}_{favor}$	$\mathbf{F}_{against}$	$\mathbf{F}_{avg}$
Majority class	52.01	78.44	65.22
SVM-unigrams	54.49	72.13	63.31
SVM-ngrams	62.98	74.98	68.98
SVM-ngrams-comb	54.11	70.01	62.06
Équipe			
MITRE	59.32	76.33	67.82
pkudblab	61.98	72.67	67.33
TakeLab	60.93	72.67	66.83

TABLE 2.1 Résultats pour la sous-tâche A des baselines et des 3 premiers participants ordonnés

On voit qu'il est difficile de faire décoller les scores mêmes avec les meilleurs systèmes. Ceci nous donne une bonne idée de l'état de l'art en 2016. Une implémentation simple avec des n-grams bien choisis et une catégorisation aident beaucoup à avoir de bons résultats. Mais est-ce toujours possible de contextualiser les données ? Il y a un biais notoire du fait de travailler avec un micro vocabulaire d'un tweet.

Baseline	$\mathbf{F}_{favor}$	$\mathbf{F}_{against}$	$\mathbf{F}_{avg}$
Majority class	0.00	59.44	29.72
SVM-ngrams-comb	18.42	38.45	28.43
Équipe			
pkudblab	57.39	55.17	56.28
LitisMind	30.04	59.28	44.66
INF-UFRGS	32.56	52.09	42.32

TABLE 2.2 Résultats pour la sous-tâche B des baselines et 3 premiers participants ordonnés

La tâche dite non-supervisée a fini par être semi-supervisée. Au final, les équipes ont utilisé le corpus d'Hillary Clinton pour former celui de Trump. Ils ont juste ajouté quelques expressions clichées pour trouver la faveur ou la défaveur du tweet. Le sujet de cette sous-tâche est particulièrement clivant. En somme, cela ne permet pas de retirer une méthode d'apprentissage non-supervisée objective.

2.3 *Fake News Challenge*

Paradoxalement, la *Fake News Challenge* (FNC) n'est pas un classificateur de véracité de news. Nous ne collons pas automatiquement un label *Fake News* ou *True News*. Si nous avons présenté la détection de parti pris jusqu'à présent c'est que la première tâche du FNC l'a utilisé. Dans le troisième chapitre de ce mémoire, nous prolongerons même son utilisation. Mais alors comment justifie-t-on l'utilisation d'une telle technique ? Et en quoi est-elle utile pour la détection de *Fake News* ?

Premièrement, à l'heure actuelle, nous essayons de faire progresser la recherche humaine dans la vérifications de faits (*Fact-checker*). Nous proposons donc de récolter rapidement les informations nécessaires pour traiter la *News* comme *Fake* ou *True*. Ici, spécifiquement, nous essayons de fournir à l'annotateur humain de récupérer instantanément les articles qui sont en accord, en désaccord ou qui discutent d'une affirmation particulière. L'annotateur humain pourrait alors examiner les arguments pour et contre une affirmation donnée. La relation étant connue, il pourrait établir la vérité ou la fausseté de l'affirmation. Le but premier est donc ici le gain de temps de recherche.

Deuxièmement, dans le futur, nous pouvons imaginer un classificateur de *Fake News*/*True News* qui se base sur un détecteur de prise de parti. Mais cela nécessiterait un corpus conséquent de la presse qui soit préalablement admis comme vrai par des annotateurs humains et un autre avec les informations classées comme fausses, aussi annoté par les humains. Entre autres, ce classificateur pourrait annoter les affirmations comme vraies ou fausses en fonction de tel ou tel corpus de données. Ainsi, si une affirmation est en accord avec un des corpus de textes contenant des informations vraies, alors on pourra dire qu'elle est vraie elle aussi. L'affirmation « la terre est un globe » serait validée pour le corpus des énoncées de la physique actuelle. Si cette même information est invalidée par le corpus de fausses informations, cela corroborera sa véracité. L'affirmation « la terre est un globe » serait invalidée pour le corpus des énoncées de la Flat Earth Society [5]. À l'inverse si une affirmation est en accord avec le corpus de fausses informations et en désaccord avec celui de bonnes informations, alors l'affirmation pourra être présentement considérée comme fausse. « La terre est plate » serait ce type d'affirmation fausse. En effet, il faudra qu'un tel classificateur puisse arranger ses corpus en fonction des preuves qu'on lui apporte sur notre réalité. Une chose fausse hier peut être vraie le lendemain, et vice-versa. « La terre a la forme d'un patatoïde complexe proche d'un globe » remplacera l'affirmation « la terre est un globe » , car elle est plus exacte que celle-ci. « La terre est un globe » ne serait pas une affirmation fausse pour autant mais une affirmation moins vraie. Dans le cas où une affirmation est en accord avec le corpus d'informations fausses et le corpus d'informations vraies, alors on pourrait établir un pourcentage de crédibilité de cette information. Par exemple, « Que la terre soit un globe ou un disque, elle tourne autour du soleil » . Et dans un dernier cas où l'affirmation ne serait en accord avec aucun des deux corpus, alors on pourrait dire que cette affirmation n'est pas liée au sujet de ces corpus ou que l'on suspend pour le moment notre jugement sur celle-ci. Le but second ici serait de hiérarchiser rapidement les informations, ce qui devrait être présent dans un moteur de recherche, par exemple.

5. Organisation soutenant l'idée que la Terre est plate.

2.3.1 Description générale de la tâche

Nous présentons dans cette section la tâche partagée qu'est le FNC et les différentes solutions proposées par ses participants. Pour détecter des *Fake News* il nous faut résoudre plusieurs défis, à savoir :

1. Déterminer si les faits présents dans l'article de presse sont corrects ; c'est-à-dire déterminer la véracité des faits par rapport à la réalité.
2. Analyser les relations entre le titre de l'article et le corps de l'article.
3. Quantifier le biais inhérent d'un texte.

On voit clairement une application de la détection de parti pris dans le défi n° 2. En effet, ce défi correspond parfaitement à sa définition.

Évaluer la véracité d'un article est une tâche complexe et lourde, même pour des experts formés. La première étape du FNC se concentre sur la tâche de détection de parti pris.

But

L'objectif du FNC est d'explorer comment les technologies d'intelligence artificielle pourraient être utilisées pour lutter contre les *Fake News*.

Organisation générale

Vous pouvez retrouver toutes les informations du *Fake News Challenge* sur leur site officiel. Les codes de leur baseline et les données sont en libres accès sur github.

Données et origines des données

Le FNC est une tâche partagée supervisée. Les données sont fournies par les organisateurs. Les organisateurs définissent les données en termes d'entrées (**C** et **E**) et de sorties (**R**). Une entrée est un titre d'article et un corps d'article, soit à partir d'un même article ou de deux articles différents. Une entrée est donc formée par le couple d'une affirmation **E** (titre d'un article) et un corps de texte **C** (corps d'un article). Une sortie est la relation **R** corps du texte par rapport à la revendication faite dans l'affirmation définie par l'une de ces quatre catégories :

related : Le corps du texte **C** et l'affirmation **E** traitent d'un sujet en commun [6].

- **agree :** Le corps du texte **C** est en accord avec l'affirmation **E**.
- **disagree :** Le corps du texte **C** n'est pas d'accord avec l'affirmation **E**.
- **discuss :** Le corps du texte discute **C** le même sujet que l'affirmation **E**, mais ne prend pas de parti pris.

unrelated : Le corps du texte **C** traite d'un sujet différent de l'affirmation **E**.

6. Cette méta-relation ne labellisera jamais les données : elle est là pour mieux comprendre le but de la détection de prise de parti.

Ferreira & Vlachos (2016) ont au préalable testé et créé un ensemble de données à partir du site du projet Emergent[11] de Craig Silverman. Le projet Emergent fut aussi utile pour la création du corpus de la FNC.

Les données de ce projet ont été collectées par des journalistes du *Tow Center for Digital Journalism*. Pour ajouter une entrée au site, le journaliste doit trouver deux choses : une affirmation qui semble être une rumeur et un ensemble d'articles qui parlent de cette soi-disant rumeur. Les sujets de chaque affirmation varient, cela va des déclarations politiques de Donald Trump aux comparaisons de prix de la prochaine Apple Watch. Les sources de celles-ci sont les comptes Twitter polémiques, traitant les rumeurs et les sites tel que Snopes.com [7]. À partir de certaines affirmations, le journaliste constitue donc un corpus d'articles pour établir la véracité de celles-ci. Chaque affirmation peut-être « Vraie » , « Fausse » ou « Non-vérifiée » . Cette valeur de vérité peut-être établie grâce au corpus d'articles (Ensemble de **C**) ; où chaque article peut soit être « Pour » , « Contre » ou « Neutre » . Le journaliste résume l'article en un gros titre (nos affirmations **E**). La véracité des affirmations **E** n'est pas basée sur le nombre d'articles « Pour » ou « Contre » ; mais bien sur la vraisemblance des preuves qui sont rapportées dans les articles jugés par le journaliste.

Le corpus de Ferreira & Vlachos est composé alors de 300 affirmations de rumeurs dont 2595 articles sont associés à un ratio de 8,75 articles pour une affirmation. La distribution hétérogène est de 47.7% « Pour » , 15.2% « Contre » et 37.1% « Neutre » . Pour tester si les données d'Emergent étaient assez discriminatoires et robustes pour une tâche d'apprentissage automatique, Ferreira & Vlachos ont testé l'accord entre les gros titres des journalistes et les affirmations de rumeurs. Nous ne détaillerons pas plus leur protocole ici. Mais leur 73% d'exactitude (par rapport aux 47% de leur baseline naïve) donne une bonne estimation de la consistance des données.

La répartitions des données de la FNC est faite avec le même procédé mais cette fois-ci on a extrait l'affirmation de rumeurs et le corps des articles. De plus, les organisateurs ajoutent une dimension de classification ; la classe **unrelated** est formée à partir des couples d'affirmation et de valeurs qui n'ont pas de liens sur la plate-forme Emergent. Les sujets sont alors non-liés. La répartition des données au niveau des relations **R** se fait donc ainsi pour l'entraînement : 73.13% sont **unrelated**, 17.82% sont **discuss**, 7.36% sont **agree** et 1.68% sont **disagree**. En chiffres, cela donne 37727 corps d'articles **C** pour 1648 affirmations de rumeurs **E**. Ce qui a donné 49973 couples **C/E**. L'ensemble de test est composé de 20019 corps d'articles **C**, 894 affirmations de rumeurs **E** et donc de 25419 couples **C/E**. Bien sûr, les données **C** ou **E** entre le test et l'entraînement ne se recoupent pas et les relations **R** ne sont pas données dans le test.

Voici un exemple de données pour l'affirmation **E** suivante « *Robert Plant Ripped up $800M Led Zeppelin Reunion Contract* » :

- **R : Agree E** : « … *Led Zeppelin's Robert Plant turned down £500 MILLION to reform supergroup.* … »
- **R : Disagree E** : « … *No, Robert Plant did not rip up an $800 million deal to get Led Zeppelin back together.* … »

7. Snopes.com est un site Web anglophone créé dans le but de limiter la propagation de canulars informatiques et de rumeurs infondées circulant sur Internet.

R : Discusses **E** : « … *Robert Plant reportedly tore up an $800 million Led Zeppelin reunion deal.* … »

R : Unrelated **E** : « … *Richard Branson's Virgin Galactic is set to launch SpaceShipTwo today.* … »

L'évaluation

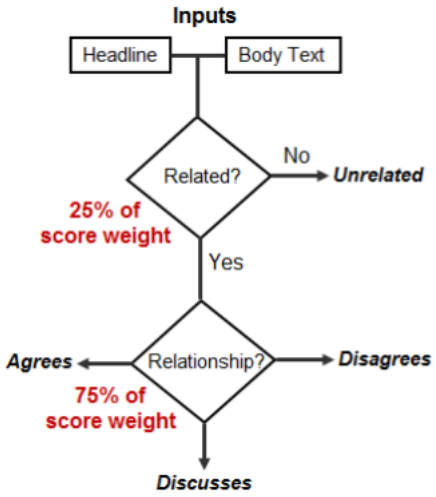


FIGURE 2.3 Diagramme de calcul du score relatif. Traduire ici « Headline » par **C** et « Body Text » par **E**

Les équipes seront évaluées selon un système de score relatif pondéré à deux niveaux ; pour le premier niveau il faut classer **E** et **C** comme étant liés ou non (25% de la pondération). Puis pour le deuxième niveau il faut classer les couples liés comme étant d'accord, en désaccord ou discuté (75% de la pondération), trouver le bon **R** du couple. Le score relatif est le score brut normalisé par le score maximum possible sur l'ensemble de test [8]. En effet, il est très facile d'avoir une très bonne exactitude avec une classe sur-représentée comme les **unrelated** (73.13% du corpus d'entraînement). Il y a donc deux règles qui régissent ce score :

unrelated/related : si la classe Gold et la classe attribuée ont la même méta-classe alors on ajoute 0.25 au score.

same related : si entre deux classes **related** la classe Gold et la classe attiribuée sont les mêmes, alors on ajoute 0.75 au score.

Exemples de scores en fonction de l'attribution de classe :

Classe Gold	Classe attribuée	Score
unrelated	unrelated	+0.25
agree	unrelated	+0
agree	disagree	+0.25
agree	agree	+0.75

8. Nous donnerons à chaque fois le score brut et le score normalisé en pourcent.

TABLE 2.3 Score en fonction de l'attribution de classe

Baseline

Une simple baseline utilisant un classificateur de *booster* de gradient est fourni par les organisateurs. Cette baseline inclut également le code pour le pré-traitement du texte, la division des données pour éviter des pertes entre l'entraînement et le test, la validation croisée avec k-fold. Cette baseline permet les overlaps entre les mots et les n-grams et des fonctions d'indicateurs pour la polarité et la réfutation. Avec ces caractéristiques et un classificateur *boosté*, la baseline atteint un score d'exactitude pondérée de 79,53% avec dix validations croisées.

Les participants

Il y a eu 71 équipes participantes pour cette tâche. Les trois premières équipes avaient pour obligation d'écrire un article sur leur système et d'en publier une version Opensource. Dans les sous-sections qui suivent, nous allons justement vous présenter les systèmes des trois gagnants du FNC.

2.3.2 Solat in the Swen[5]

Méthodologie

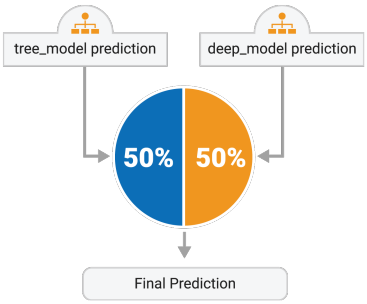


FIGURE 2.4 Modèle avec deux sous-modèles concurrents de Solat[5]

L'équipe de Solat in the Swen a implémenté plusieurs modèles performants. Ils ont ensuite décidé de faire un modèle utilisant des systèmes concurrents. Les traits entre la solution de Solat et la baseline de la FNC ne changent pas. Ils utilisent le même *pre-processing* et la même vectorisation. Le modèle est composé de plusieurs systèmes concurrents : le premier est un *deep convolutional neural network* [9] et le second est un *gradient-boosted decision trees* [10] [11].

Détaillons ici les algorithmes du modèle :

9. Dans l'apprentissage automatique, un réseau de neurones convolutionnels est une classe de réseaux neuronaux artificiels profonds qui a souvent été appliquée avec succès à l'analyse de l'imagerie visuelle. Un réseau de neurones convolutionnels utilise une variation de perceptrons multi-couches conçus pour nécessiter un prétraitement minimal.

10. Un arbre de décision est un outil d'aide à la décision qui utilise un graphe arborescent ou un modèle de décisions et leurs conséquences possibles, y compris les résultats d'événements aléatoires, les coûts des ressources et l'utilité.

11. Le *boosting* de gradient est une technique d'apprentissage automatique pour les problèmes de régression et de classification, qui produit un modèle de prédiction sous la forme d'un ensemble de modèles de prédiction comme pour des arbres de décision.

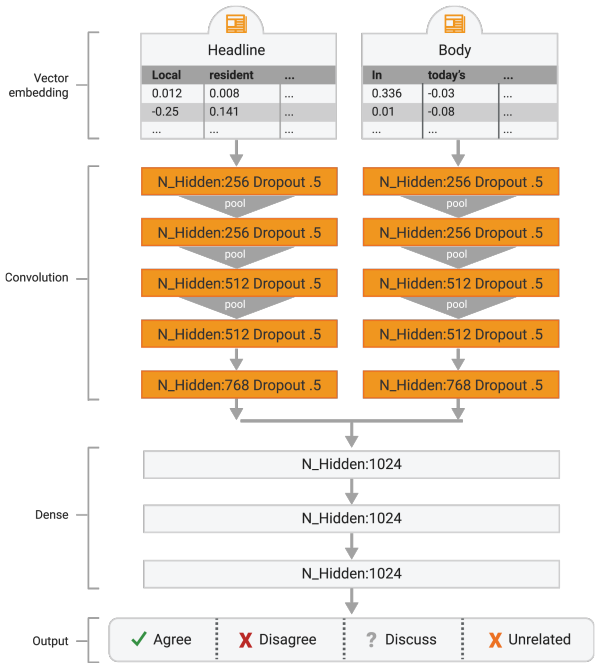


FIGURE 2.5 Approche d'apprentissage en profondeur ; modèle CNN + Multi layer perceptron[5]

Le premier modèle utilisé par l'équipe a plusieurs réseaux de neurones différents utilisés dans l'apprentissage en profondeur. Ce modèle s'applique à la convolution numérique nette unidimensionnelle (CNN) sur le titre et le corps du texte, en utilisant les vecteurs préchargés de Google News. Les CNN permettent un calcul parallèle efficace lors de l'exécution. La sortie de ce CNN est ensuite envoyée à MLP avec une sortie 4 classes : « agree » , « disagree » , « discuss » et « unrelated » , formées de bout en bout. L'architecture de ce modèle a été choisie en raison de sa facilité de mise en œuvre et de son calcul rapide, puisque l'on peut compter sur des convolutions au lieu de récurrence. Ce modèle est toutefois limité par le fait qu'il ne peut observer le texte qu'une seule fois.

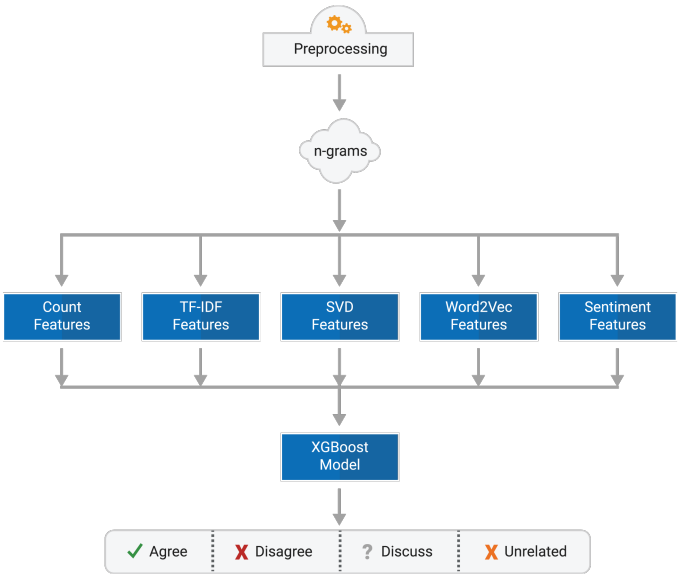


FIGURE 2.6 Approche Gradient-Boosted Decision Trees[5]

L'autre modèle utilisé dans l'ensemble est un modèle d'arbres de décision à gradient d'intensité (GBDT). Ce modèle introduit peu de caractéristiques textuelles de l'affirmation de rumeurs et du corps de l'article, qui sont ensuite introduites dans le sujet d'un gradient dans la relation entre **C** et **E**. Ce modèle basé sur le texte entraîné avec XGBoost utilise plusieurs modules de décision à savoir :

Un compteur de trait pour compter le nombre de traits communs entre le titre et le corps de l'article.

La méthode de pondération TF-IDF pour comparer de manière inversée la fréquence relative des mots.

L'algèbre linéaire SVD Features pour mesurer la similarité entre différents n-grams.

L'espace vectoriel word2vec pour comparer le cosinus de similarité entre deux représentations vectorielles.

Un module Sentiment utilisant un corpus de mots connotés sentimentalement.

Les arbres de décision à gradient ont été choisis en raison de la robustesse du modèle par rapport aux différentes échelles des vecteurs caractéristiques.

Résultats

Solat in the Swen gagne le FNC avec un score de 9556.50 points donc un score relatif de 82.02%.

2.3.3 Athene (UKP Lab)[15]

Méthodologie

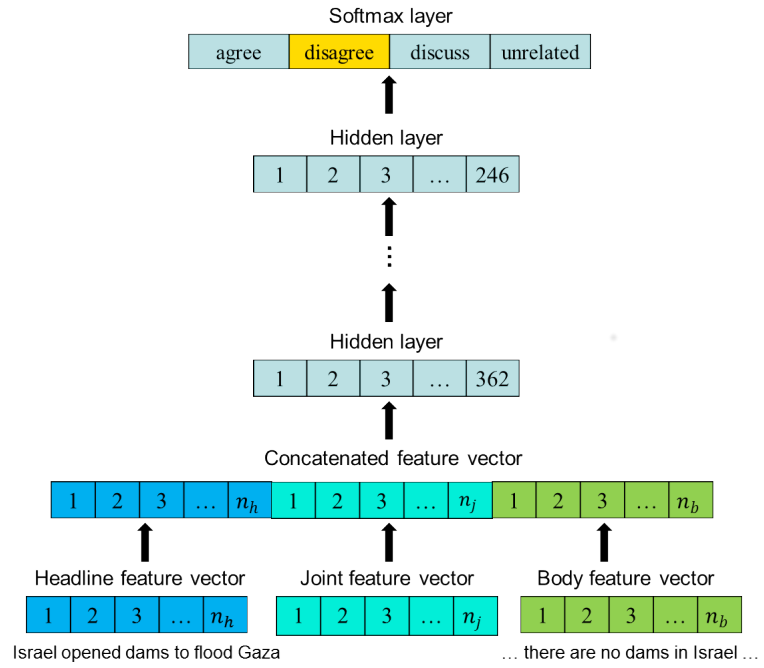


FIGURE 2.7 *Multi layer perceptron* utilisé pour le FNC par l'équipe Athene[15]

Le *Multi layer perceptron* proposé par Richard Davis et Chris Proctor (organisateurs du FNC) a été le point de départ pour le développement du système final d'Athene. La structure du système a été optimisée pour les caractéristiques par une recherche aléatoire par laquelle les hyper-paramètres ont été ajustés. La structure du modèle qui en résulte est illustrée ci-dessus en résumé, car 5 des 7 couches cachées sont ignorées.

Le pre-processing d'Athene utilise différents traits, à savoir : les Bag of Words sur les uni-grams, les matrices de factorisations non-négatives, l'indexation et l'analyse latente sémantique (LSA, LSI) et la détection de paraphrase basée sur le recouvrement de mots (avec word2vec).

La vectorisation des informations se passe en trois étapes. La vectorisation des données pré-processées de **C** puis de **E** et la création d'un vecteur joint des dimensions se recoupant dans les deux précédents vecteurs. Ces trois vecteurs sont alors concaténés en un seul pour former une entrée.

Le modèle final rassemble cinq *Multi layer perceptron* initialisés de manière aléatoire, qui donnent leurs sorties à un seul *Multi layer perceptron* qui va faire le vote de la classes à attribuer à partir des autres *Multi layer perceptron*.

Résultats

Athene arrive deuxième au classement du FNC avec un score de 9550.75 points, donc un score relatif de 81.97%.

2.3.4 UCL Machine reading[22]

Méthodologie

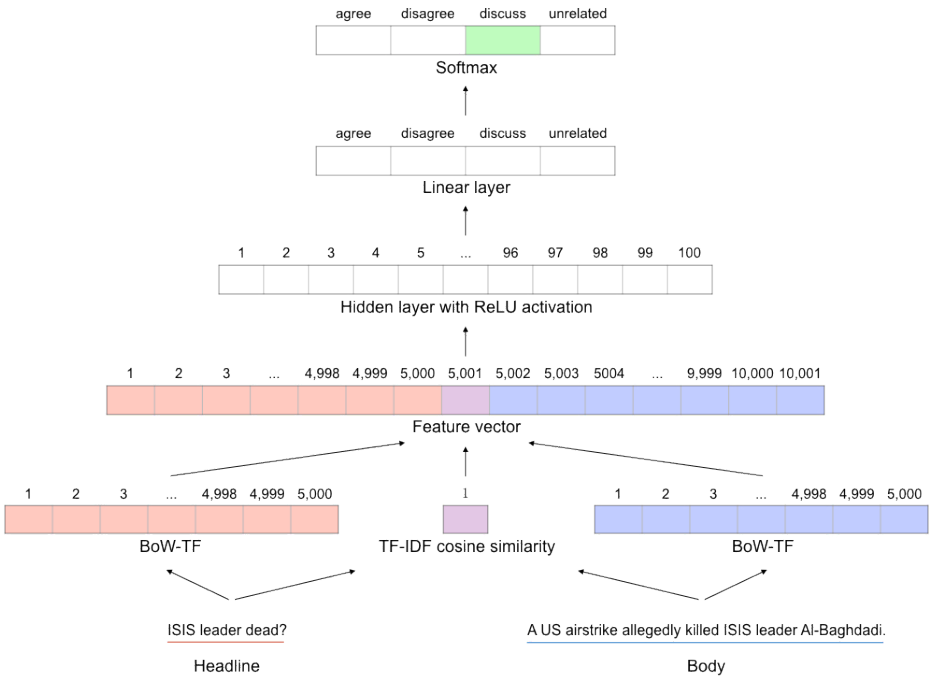


FIGURE 2.8 Schéma du modèle de l'UCL-MR[22]

UCL-MR est juste un *Multi layer perceptron* (plutôt simple par rapport aux autres solutions) entraîné sur les unigrams et une utilisation astucieuse du TF-IDF.

UCL utilise deux représentations simples des mots avec BoW pour les entrées de texte : fréquence de terme (TF) pour représenter l'affirmation **C** et le corps de l'article **E** et l'inverse de la fréquence des mots du document (TF-IDF) pour calculer le cosinus de similarité entre **E** et **C**.

Ainsi les vecteurs d'entrée contiennent ces trois vecteurs composites. Les vecteurs passent dans une seule couche d'entrée de 100 perceptrons avec une activation ReLU. Puis une couche liénaire pour donner une valeur à chacune des classes. Puis un softmax pour sortir la classe dominante.

UCL-MR utilise juste un *Multi layer perceptron* mais avec des paramètres très optimisés. Nous ne détaillerons pas ici les paramètres d'implémentation qui se retrouvent dans leur publication.

Résultats

UCL-MR arrive troisième au classement du FNC avec un score de 9521.50 points donc un score relatif de 81.72%.

2.3.5 Discussion comparative des modèles et des résultats

Équipes	Scores bruts	Scores relatifs
Solat in the Swen	9556.50	82.02%
Athene	9550.75	81.97%
UCL-mr	9521.50	81.72%

TABLE 2.4 Résultats ordonnés de la FNC

Comme l'a dit Dean Domerleau (organisateur) en parlant de ce tableau de résultats : « Ce que tout cela signifie, c'est que les meilleures équipes se sont très bien débrouillées, mais il y a certainement encore de la place pour l'amélioration ! » . En effet, on pourrait pousser plus loin les systèmes pour que ceux-ci nous donnent de meilleurs résultats. Mais le fait qu'il n'y ait pas de percée d'une équipe dans le tableau des scores, et le fait que très peu d'équipes aient fait mieux que la baseline, montrent que cette tâche est difficile.

De plus, un système complexe comme celui de Solat, censé être explicatif sur les traits pertinents, ne se démarque que de 0,3 points du système d'UCL-MR, très peu explicatif mais bien optimisé, qui reste une boîte noire au niveau de la sélection des paramètres. On revient à un modèle régressif qui est simple et qui explique beaucoup avec peu. Mais cela donne une compréhension minime du phénomène. Ou peut-être le système d'UCL-mr est surévalué pour cette ensemble de données. En ce qui concerne Athene, les résultats sont sensiblement les mêmes que Solat. La complexité de leur modèle est aussi faite à partir d'apprentissages automatisés, certes moins complexe que Solat, mais qui reste quand même un modèle avec des systèmes concurrents.

Ce qu'il faudrait faire, c'est comparer les outputs de chaque système et voir comment ils trient les classes **related**. Ainsi on pourrait faire des recoupements sur les données. Cela permettrait de constater ce que les différents systèmes font à l'identique et ce qu'ils font différemment. On déterminerait alors les traits et les architectures de modèle relatifs aux classes de données.

C'est ce que je me propose de faire dans la première section de la partie 3 de ce mémoire, dans le but ensuite d'apporter une contribution originale à ce problème en partant des systèmes déjà réalisés.

Chapitre 3

Recherches et résultats

3.1 Analyse des résultats des participants et création d'hypothèses

Dans cette troisième partie nous tentons de faire une analyse comparative d'erreurs des participants de la FNC. Nous avons reproduit les résultats de tous les participants.

3.1.1 Un aperçu de l'ensemble de test

Il y a 7064 points à marquer avec les *Related* et seulement 4587.25 points avec les *Unrelated*. Ainsi, comme nous l'avions vu précédemment, trouver un *Related* rapporte 4 fois plus de points que trouver un *Unrelated*.

Afin de produire des hypothèses non-biaisées [1], nous allons faire des observations des résultats des différents modèles sur 80% des entrées de l'ensemble de test [2].

Stance	unrelated	agree	disagree	discuss	Somme
Nombre	14662	1568	550	3550	20330
Pourcentage	72.12%	7.71%	2.70%	17.46%	100%
Score FNC	3665.5	1568	550	3550	9333.5

TABLE 3.1 Répartition des données dans l'ensemble de test à 80%

Le corpus de test étant partiellement ordonné, nous avons retiré les 20% de manière ordonnée [3] aussi.

1. Du moins, des hypothèses qui ne soient pas ajustées totalement à l'ensemble de test final.
2. Retrouvez en annexe les résultats complets de chaque modèle gagnant. Ces résultats seront utilisés pour comparer les résultats de nos propres modèles. En revanche, comme précisé, ils ne seront pas utilisés pour formuler des hypothèses.
3. En effet, nous avons retiré toutes les entrées dont l'indice était divisible par 5 ou 10

3.2 Les résultats et les scores des participants

Ici, nous présentons les différents résultats des modèles de chaque participant sur notre ensemble de test à 80%. Notez que les tables de confusion ont horizontalement les labels de vérité et verticalement les prédictions. Nous présentons les participants du premier au dernier selon le score FNC.

3.2.1 Solat in the Swen

Solat Complet

	agree	disagree	discuss	unrelated	Somme
agree	927	11	478	152	1568
disagree	217	8	235	90	550
discuss	661	5	2700	184	3550
unrelated	26	0	172	14464	14662
Somme	1831	24	3585	14890	20330

TABLE 3.2 Table de confusion du modèle : Solat Complet

On remarque que la classe **disagree** est largement sous-représentée. Cette classe est aussi sous-représentée dans l'ensemble d'entraînement. Ce qui explique potentiellement pourquoi nous ne détectons pas ces traits distinctifs.

Solat Complet

Mesure	agree	disagree	discuss	unrelated
Précision	0.59	0.01	0.76	0.99
Rappel	0.51	0.33	0.75	0.97
F1score	0.55	0.03	0.76	0.98
Exactitude	89.03			
Score FNC	7652.75			
Pourcentage FNC	81.99			

TABLE 3.3 Mesures pour le modèle : Solat Complet

Bien que Solat in the Swen ait remporté la 1ère place de la FNC, l'exactitude et les scores ci-dessus ne sont pas les maximaux.

Les sous-modèles de Solat in the Swen

Sous-modèle arborescent de Solat in the Swen Les sous-modèles de Solat n'ont pas eu de soumission à la FNC mais ils sont quand même intéressants à étudier car ils expliquent les résultats et les biais du modèle.

Solat Arborescent

	agree	disagree	discuss	unrelated	Somme
agree	807	0	690	71	1568
disagree	147	1	334	68	550
discuss	500	1	2902	147	3550
unrelated	16	0	160	14486	14662
Somme	1470	2	4086	14772	20330

TABLE 3.4 Table de confusion du modèle : Solat Arborescent

Nous pouvons observer ici que le modèle principal a une aversion pour le label **disagree**.

Solat Arborescent

Mesure	agree	disagree	discuss	unrelated
Précision	0.51	0.0	0.82	0.99
Rappel	0.55	0.5	0.71	0.98
F1score	0.53	0.0	0.76	0.98
Exactitude	89.5			
Score FNC	7749.5			
Pourcentage FNC	83.03			

TABLE 3.5 Mesures pour le modèle : Solat Arborescent

Ce sous-modèle ne tient pas vraiment compte de la classe **disagree** mais il a paradoxalement tout de même les plus hauts scores du challenge. En effet, si Solat avait proposé uniquement ce modèle il aurait pris encore plus d'avance par rapport aux autres participants.

Sous-modèle de Deep Learning de Solat in the Swen

Solat DeepLearning

	agree	disagree	discuss	unrelated	Somme
agree	911	115	143	399	1568
disagree	271	62	41	176	550
discuss	1396	137	1397	620	3550
unrelated	2321	449	766	11126	14662
Somme	4899	763	2347	12321	20330

TABLE 3.6 Table de confusion du modèle : Solat DeepLearning

Nous voyons clairement que ce sous-modèle vient équilibrer le sous-modèle arborescent. Sa table de confusion montre une tentative d'uniformisation de la classe **disagree**.

Solat deep

Mesure	agree	disagree	discuss	unrelated
Précision	0.58	0.11	0.39	0.76
Rappel	0.19	0.08	0.6	0.9
F1score	0.28	0.09	0.47	0.82
Exactitude	66.38			
Score FNC	5677.25			
Pourcentage FNC	60.83			

TABLE 3.7 Mesures pour le modèle : Solat DeepLearning

Cette uniformisation a pour conséquence que ce sous-modèle est le pire de tous les modèles. Combiné avec le sous-modèle arborescent, il permet d'avoir une augmentation minime de la classe **disagree** au détriment des autres classes.

3.2.2 Le système Athene

Athene

	agree	disagree	discuss	unrelated	Somme
agree	709	57	669	133	1568
disagree	193	56	193	108	550
discuss	388	30	2853	279	3550
unrelated	16	3	86	14557	14662
Somme	1306	146	3801	15077	20330

TABLE 3.8 Table de confusion du modèle : Athene

La table de confusion d'Athene ressemble beaucoup à celle de Solat. Ce modèle tente beaucoup plus de labelliser des titres d'articles comme **disagree**. Mais il a beaucoup plus de mal à distinguer la classe **discuss** de la classe **agree**.

Athene

Mesure	agree	disagree	discuss	unrelated
Précision	0.45	0.1	0.8	0.99
Rappel	0.54	0.38	0.75	0.97
F1score	0.49	0.16	0.78	0.98
Exactitude	89.4			
Score FNC	7639.75			
Pourcentage FNC	81.85			

TABLE 3.9 Mesures pour le modèle : Athene

Athene a la meilleure exactitude de toute la FNC car elle classe très bien les **unrelated** (qui ne valent pas beaucoup de points dans le score FNC).

3.2.3 UCL Machine Reader

UCL-mr					
	agree	disagree	discuss	unrelated	Somme
agree	697	7	770	94	1568
disagree	144	37	277	92	550
discuss	424	35	2882	209	3550
unrelated	44	2	261	14355	14662
Somme	1309	81	4190	14750	20330

TABLE 3.10 Table de confusion du modèle : UCL-mr

Nous observons une tentative d'uniformisation proportionnelle de la table de confusion. UCL-mr sait très bien classer les **unrelated**. Néanmoins, il manque apparemment de traits pour distinguer les **agree** des **discuss**.

UCL-mr				
Mesure	agree	disagree	discuss	unrelated
Précision	0.44	0.07	0.81	0.98
Rappel	0.53	0.46	0.69	0.97
F1score	0.48	0.12	0.74	0.98
Exactitude	88.4			
Score FNC	7619.0			
Pourcentage FNC	81.63			

TABLE 3.11 Mesures pour le modèle : UCL-mr

3.3 Analyses et hypothèses

Voici une liste d'hypothèses basée sur nos constatations que nous tenterons de vérifier par la suite avec les différents modèles. Nous voyons clairement une confusion entre la classe **agree** et la classe **discuss**. Mieux distinguer ces deux classes rapporterait des points. Cela pourrait faire une différence significative. La classe **disagree** est sous-représentée donc quantitativement, elle ne rapporte que peu de points. Le modèle de Solat semble plus robuste que les autres en utilisant l'apprentissage par ensemble (*ensemble learning*). Utiliser ce type d'apprentissage entre les modèles des participants augmenterait certainement les performances [4].

4. Cette idée d'apprentissage par ensemble a été de nombreuses fois démontrée comme efficace comme dans Opitz, D.; Maclin, R. (1999). « Popular ensemble methods: An empirical study » . Mais nous ne développerons pas plus ici l'état de l'art de cette technique.

3.4 Modèles par combinaison : le vote de majorité

3.4.1 Explication du modèle

Comme nous l'avons vu ci-dessus, les combinaisons entre les modèles déjà présentés à la FNC peuvent atteindre de hauts résultats. Notre premier modèle sera un jet naïf de combinaisons de deux modèles. Le vote de majorité se base sur les classes prédites par les modèles des participants pour l'ensemble de test complet. Ce vote de majorité marche avec un modèle dominant départageant en cas de conflit entre les deux autres modèles assujettis. La classe la plus votée sera la classe prédite. Si jamais les trois classes sont toutes différentes, alors la classe donnée par le modèle dominant est choisie.

Vote de majorité entre 3 modèles

Modèle Dominant	Modèle assujetti 1	Modèle assujetti 2	Classe prédite
agree	**agree**	unrelated	**agree**
disagree	**discuss**	**discuss**	**discuss**
discuss	disagree	agree	**discuss**

TABLE 3.12 Exemple du vote de majorité

Ainsi, chaque modèle à tour de rôle peut devenir le modèle dominant, ce qui donne les résultats dans la sous-section suivante.

3.4.2 Les résultats des votes de majorité

Dans la suite de ce chapitre, nous allons voir beaucoup de résultats de nos modèles. Pour se donner un référentiel, nous utiliserons comme baseline les résultats de Solat complet [5]. Nous comparerons aussi nos modèles avec leurs modèles parents, quand cela est pertinent.

Dominant : Solat in the Swen

Vote de majorité avec dominant : Solat

	agree	disagree	discuss	unrelated	Somme
agree	923	13	815	152	1903
disagree	228	20	321	128	697
discuss	483	4	3729	248	4464
unrelated	27	0	149	18173	18349
Somme	1661	37	5014	18701	25413

TABLE 3.13 Table de confusion du modèle : Vote de majorité Solat

Par rapport au Solat complet : les **agree** chutent de 191 points, les **disagree** gagnent 7 points, les **discuss** gagnent 328 points et les **unrelated** gagnent 238 points.

5. Vous pouvez retrouver ces résultats dans les annexes A.

Vote de majorité avec dominant : Solat

Mesure	agree	disagree	discuss	unrelated
Précision	0.49	0.03	0.84	0.99
Rappel	0.56	0.54	0.74	0.97
F1score	0.52	0.05	0.79	0.98
Exactitude	89.89			
Score FNC	9681.25			
Pourcentage FNC	83.09			

TABLE 3.14 Mesures pour le modèle : Vote de majorité Solat

Quand Solat est dominant, on gagne 1% au score FNC. Cela est dû au renfort des autres modèles sur les labels **disagree**, **discuss** et **unrelated**. Par contre, nous notons que Solat est meilleur individuellement sur le label **agree**.

Dominant : Athene

Vote de majorité avec dominant : Athene

	agree	disagree	discuss	unrelated	Somme
agree	914	42	809	138	1903
disagree	213	46	308	130	697
discuss	453	21	3735	255	4464
unrelated	14	2	148	18185	18349
Somme	1594	111	5000	18708	25413

TABLE 3.15 Table de confusion du modèle : Vote de majorité Athene

Par rapport au Solat complet : les **agree** chutent de 200 points, les **disagree** gagnent 33 points, les **discuss** gagnent 334 points et les **unrelated** gagnent 74 points.

Par rapport au Athene complet : les **agree** gagnent 63 points, les **disagree** chutent de 20 points, les **discuss** gagnent 124 points et les **unrelated** chutent de 26 points.

Vote de majorité avec dominant : Athene

Mesure	agree	disagree	discuss	unrelated
Précision	0.48	0.07	0.84	0.99
Rappel	0.57	0.41	0.75	0.97
F1score	0.52	0.11	0.79	0.98
Exactitude	90.03			
Score FNC	9702.75			
Pourcentage FNC	83.28			

TABLE 3.16 Mesures pour le modèle : Vote de majorité Athene

Dominant : UCL Machine Reader

Quand Athene est le modèle dominant, les pertes sont vraiment minimes. Ce modèle profite du gain significatif d'Athene sur le label **discuss**.

Vote de majorité avec dominant : UCL-mr

	agree	disagree	discuss	unrelated	Somme
agree	920	13	844	126	1903
disagree	202	35	339	121	697
discuss	467	26	3728	243	4464
unrelated	19	1	157	18172	18349
Somme	1608	75	5068	18662	25413

TABLE 3.17 Table de confusion du modèle : Vote de majorité UCL-mr

Par rapport au Solat complet : les **agree** chutent de 194 points, les **disagree** gagnent 22 points, les **discuss** gagnent 327 points et les **unrelated** gagnent 61 points.

Par rapport au UCL-mr complet : les **agree** gagnent 82 points, les **disagree** chutent de 11 points, les **discuss** chutent de 95 points et les **unrelated** chutent de 209 points.

Vote de majorité avec dominant : UCL-mr

Mesure	agree	disagree	discuss	unrelated
Précision	0.48	0.05	0.84	0.99
Rappel	0.57	0.47	0.74	0.97
F1score	0.52	0.09	0.78	0.98
Exactitude	89.93			
Score FNC	9698.75			
Pourcentage FNC	83.24			

TABLE 3.18 Mesures pour le modèle : Vote de majorité UCL-mr

Le modèle UCL-mr permet un très bon gain sur les autres modèles sur les **unrelated**. Malheureusement, c'est le label qui vaut le moins de points au score FNC.

Commentaires sur les résultats du vote de majorité

Bien que ce modèle d'apprentissage par ensemble semble être le plus naïf, ces résultats sont les plus hauts que l'on obtiendra. On remarque qu'Athene est le meilleur modèle en tant que dominant. C'est le modèle qui se trompe le moins en cas de conflit.

Nous remarquons aussi des spéciations des modèles en les comparant les uns avec les autres. En effet, Solat classe bien les **agree**, comparé aux autres, alors que les deux autres modèles sont bien meilleurs pour les autres labels.

3.5 Modèles par combinaison : par moyenne

3.5.1 Modèles utilisant un *50/50 weighted average*

Pour ce modèle, deux sous-modèles généreront des probabilités de label pour chaque entrée. Nous nous inspirerons de la méthode d'unification de Solat. Nous utiliserons donc aussi un *50/50 weighted average* pour combiner les résultats. Ainsi, chaque sous-modèle produira un vecteur contenant les probabilités attribuées pour chaque label (X ou Y ci-dessous). Chaque couple de mêmes dimensions (X[i] et Y[i]) passera par un *50/50 weighted average*. Le maximum des 4 résultats (max(Z)) des moyennes indiquera la classe prédite (Label) pour l'entrée.

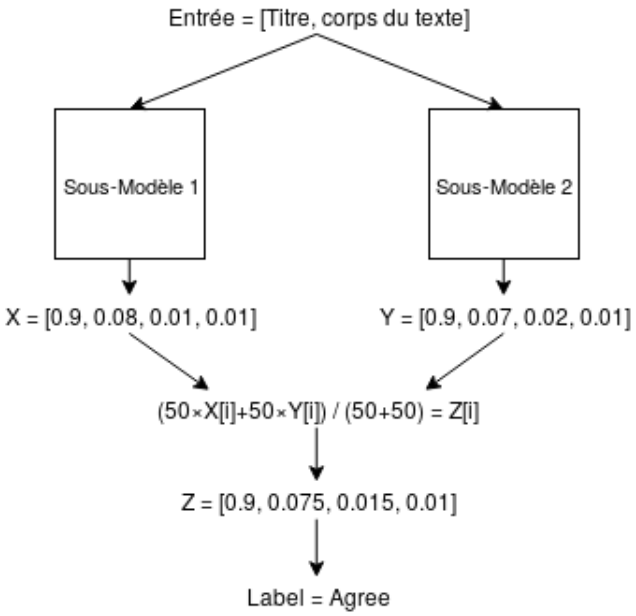


FIGURE 3.1 Diagramme explicatif du modèle *50/50 weighted average*

Notre première architecture utilisera le sous-modèle arborescent de Solat et le modèle de UCL Machine Reader. Nous appellerons ce modèle « mixte UCL-mr/Solat TF-Idf Moyenne » .

Notre deuxième architecture utilisera aussi le sous-modèle arborescent de Solat et le modèle de UCL Machine Reader mais dans cette version, le sous modèle arborescent de Solat n'utilisera pas son module de TF-Idf. La raison est que le système de UCL utilise déjà et uniquement ces traits-là. Ainsi, pour augmenter la différence entre les résultats des deux sous-modèles, nous décidons de retirer ce module. De plus, il sera entraîné seulement sur les données d'entraînement pour créer son espace sémantique [6]. Nous appellerons ce modèle « mixte UCL-mr/Solat sans TF-Idf Moyenne » .

6. En effet, dans la version originale du modèle, lors de la création de l'espace sémantique, les données de test viennent augmenter les données d'entraînement. Cela a pour conséquence que le modèle aura déjà vu les données de test lors de son entraînement. Mais ne nous méprenons pas : cela est autorisé dans le règlement de la FNC même si cela est assez discutable.

3.5.2 Résultats

Modèles mixte UCL-mr/Solat TF-Idf

Mixte UCL-mr/Solat TF-Idf

	agree	disagree	discuss	unrelated	Somme
agree	963	0	819	121	1903
disagree	198	1	377	121	697
discuss	606	3	3612	243	4464
unrelated	21	0	166	18162	18349
Somme	1788	4	4974	18647	25413

TABLE 3.19 Table de confusion du modèle : Mixte UCL-mr/Solat TF-Idf Moyenne

Par rapport au Solat complet : les **agree** chutent de 151 points, les **disagree** chutent de 12 points, les **discuss** gagnent 211 points et les **unrelated** gagnent 51 points.

Par rapport au UCL-mr complet : les **agree** gagnent de 125 points, les **disagree** chutent de 45 points, les **discuss** chutent de 21 points et les **unrelated** gagnent 199 points.

Cette table de confusion ressemble beaucoup à celle de Solat Complet. Elle montre en effet une amélioration générale des **unrelated**. Les deux modèles se complètent sur les labels **agree** et **discuss**. Cette amélioration coûte plus cher à Solat qu'à UCL-mr surtout au niveau des **agree**. Sans surprise, la classe disagree n'est pas améliorée.

Mixte UCL-mr/Solat TF-Idf

Mesure	agree	disagree	discuss	unrelated
Précision	0.51	0.0	0.81	0.99
Rappel	0.54	0.25	0.73	0.97
F1score	0.52	0.0	0.77	0.98
Exactitude	89.47			
Score FNC	9617.25			
Pourcentage FNC	82.54			

TABLE 3.20 Mesures pour le modèle : Mixte UCL-mr/Solat TF-Idf Moyenne

Nous constatons un gain de 0.52% par rapport au meilleur des systèmes de la FNC. Ceci est peu mais montre une amélioration minime de la précision de la classe **agree** et de la classe **discuss**. La classe **disagree** est encore négligée par le modèle arborescent de Solat. En conséquence, le modèle UCL-mr biaisé par Solat ne permet pas d'améliorer la précision de la classe **disagree**. Mais cela permet d'ajuster les **agree** et les **discuss**.

Modèles intermédiaires

Ici nous présentons le modèle intermédiaire « Solat arborescent Intermédiaire » afin de mieux mettre en valeur les variables indépendantes en action dans les deux autres modèles suivants [7].

7. Mixte UCL-mr/Solat sans TF-Idf Moyenne et Mixte UCL-mr/Solat sans TF-Idf SLL

Solat arborescent Intermédiaire

	agree	disagree	discuss	unrelated	Somme
agree	1021	0	785	97	1903
disagree	207	2	402	86	697
discuss	657	2	3571	234	4464
unrelated	16	0	183	18150	18349
Somme	1901	4	4941	18567	25413

TABLE 3.21 Table de confusion du modèle : Solat Intermédiaire

Par rapport au Solat complet : les **agree** chutent de 93 points, les **disagree** chutent de 11 points, les **discuss** gagnent 370 points et les **unrelated** gagnent 39 points.

Solat arborescent Intermédiaire

Mesure	agree	disagree	discuss	unrelated
Précision	0.54	0.0	0.8	0.99
Rappel	0.54	0.5	0.72	0.98
F1score	0.54	0.01	0.76	0.98
Exactitude	89.5			
Score FNC	9644.75			
Pourcentage FNC	82.78			

TABLE 3.22 Mesures pour le modèle : Solat arborescent Intermédiaire

Il ne faut pas oublier que nous avons pris comme baseline le modèle de Solat Complet. Le modèle arborescent avait déjà pour rappel les meilleurs résultats lors de notre analyse. Il va de soi qu'un modèle dont le simple entraînement diffère de quelque entrée par rapport au corpus d'entraînement original, arrive presque à des résultats similaires.

Nous observons de plus que le module TF-IDF ne doit pas être si déterminant dans le résultat final.

Le modèle D'UCL-mr Complet ne s'entraîne pas sur les données de test *(les headlines)*. Le modèle intermédiaire est similaire au modèle complet.

Modèles mixte UCL-mr/Solat sans TF-Idf

Mixte UCL-mr/Solat sans TF-Idf Moyenne

	agree	disagree	discuss	unrelated	Somme
agree	1002	0	776	125	1903
disagree	216	1	357	123	697
discuss	602	2	3596	264	4464
unrelated	16	0	149	18184	18349
Somme	1836	3	4878	18696	25413

TABLE 3.23 Table de confusion du modèle : mixte UCL-mr/Solat sans TF-Idf Moyenne

Par rapport au Solat complet : les **agree** chutent de 112 points, les **disagree** chutent de 12 points, les **discuss** gagnent 195 points et les **unrelated** gagnent 73 points.

Par rapport au UCL-mr complet : les **agree** gagnent 164 points, les **disagree** chutent de 45 points, les **discuss** chutent de 37 points et les **unrelated** gagnent 733 points.

Par rapport au modèle mixte UCL-mr/Solat TF-Idf, ce modèle améliore la classe des **agree** d'une quarantaine d'entrée. C'est certes un avantage minime, qui n'en reste pas moins surprenant. En effet, le modèle avec TF-Idf a un espace sémantique plus exhaustif et plus proche de l'ensemble de test.

Mixte UCL-mr/Solat sans TF-Idf Moyenne

Mesure	agree	disagree	discuss	unrelated
Précision	0.53	0.0	0.81	0.99
Rappel	0.55	0.33	0.74	0.97
F1score	0.54	0.0	0.77	0.98
Exactitude	89.65			
Score FNC	9633.25			
Pourcentage FNC	82.68			

TABLE 3.24 Mesures pour le modèle : mixte UCL-mr/Solat sans TF-Idf Moyenne

Les deux modèles s'ajustent l'un l'autre. UCL-mr fait gagner des **discuss** à Solat et celui-ci lui fait gagner des **agree**.

Certes, l'avantage sur le modèle précédent n'est que de 0.16% pour les pourcentage FNC. Peut-être montrons-nous qu'un espace sémantique simplement formé avec l'ensemble d'entraînement est suffisant.

3.6 Apprentissage par ensemble : *Single Layer Learner*

3.6.1 Légitimation de l'apprentissage par ensemble

Pour le moment, nous n'avons vu que des solutions qui utilisent des combinaisons de modèles. Nous n'avons pas fait à proprement parler d'apprentissage par ensemble. Pourtant, les méthodes d'ensemble utilisent des algorithmes d'apprentissage multiples pour obtenir de meilleures performances prédictives qui pourraient être obtenues à partir de l'un des algorithmes d'apprentissage constitutifs seulement[18].

Ainsi, dans cette section finale de nos solutions, nous allons utiliser une méthode d'apprentissage par ensemble qui utilise une couche neuronale apprenante à partir des résultats de deux sous-modèles.

3.6.2 Explication du modèle

Une des meilleures manières *a priori* de déterminer une classe à partir d'une liste de probabilités est certainement d'utiliser une couche neuronale. Ainsi, nous avons créé avec le framework Keras un *Single Layer Learner*. Celui-ci dispose d'une couche d'entrée de la longueur de nos vecteurs d'entraînement, d'une couche de 32 unités d'apprentissage et d'une couche de sortie de 4 unités, une par label. Afin de produire les vecteurs d'entraînement, nous avons utilisé l'intégralité du corpus d'entraînement. Chaque dixième des vecteurs a été produit par les 90% restant du corpus. Les données qui ont servi à produire les vecteurs d'entraînement sont les vecteurs concaténés du modèle Mixte UCL-mr/Solat sans TF-Idf Moyenne. Nous nommerons ce modèle Mixte UCL-mr/Solat sans TF-Idf SLL.

Code du modèle

```
import numpy as np
import keras
from keras.models import Sequential
from keras.layers import Dense, Activation
import tensorflow as tf


def create_trained_nn(vecs_train, labels_train, epochs=50):
    """Create a trained neural network model."""
    input_dim = vecs_train.shape[1]
    model = Sequential()
    model.add(Dense(units=32, input_dim=input_dim))
    model.add(Activation('relu'))
    model.add(Dense(units=4))
    model.add(Activation('softmax'))
    model.compile(loss='sparse_categorical_crossentropy',
                  optimizer='sgd',
```

```
                metrics=['accuracy'])
    model.fit(vecs_train, labels_train, epochs=epochs,
              batch_size=32)

    return model
```

3.6.3 Résultats

Mixte UCL-mr/Solat sans TF-Idf SLL

	agree	disagree	discuss	unrelated	Somme
agree	1059	3	729	112	1903
disagree	253	14	323	107	697
discuss	684	7	3530	243	4464
unrelated	45	0	288	18016	18349
Somme	2041	24	4870	18478	25413

TABLE 3.25 Table de confusion du modèle : Mixte UCL-mr/Solat sans TF-Idf SLL

Par rapport au Solat complet : les **agree** chutent de 55 points, les **disagree** gagnent un point, les **discuss** gagnent 129 points et les **unrelated** chutent de 95 points.

Par rapport au UCL-mr complet : les **agree** gagnent 191 points, les **disagree** chutent de 32 points, les **discuss** chutent de 103 points et les **unrelated** gagnent 53 points.

À chaque fois qu'un modèle chute, l'autre gagne et vice-versa. Cela ne permet pas d'améliorations.

Mixte UCL-mr/Solat sans TF-Idf SLL

Mesure	agree	disagree	discuss	unrelated
Précision	0.56	0.02	0.79	0.98
Rappel	0.52	0.58	0.72	0.97
F1score	0.54	0.04	0.76	0.98
Exactitude	89.01			
Score FNC	9606.75			
Pourcentage FNC	82.45			

TABLE 3.26 Mesures pour le modèle : Mixte UCL-mr/Solat sans TF-Idf SLL

Étonnamment, ce modèle a de moins bons résultats que le modèle précédent. Même en changeant les paramètres d'apprentissage (nombre d'unités d'apprentissage ou nombre de récursions) le modèle n'excède pas le précédent.

3.7 Discussions

Notre objectif dans ce troisième chapitre était de proposer des améliorations aux systèmes présentés à la FNC. Sur la diversité des hypothèses et des tests effectués nous observons qu'il y a très peu de systèmes qui dépassent la baseline que nous nous étions fixée (à savoir les résultats de Solat in the Swen).

Rappelons que nous avons utilisé trois méthodes de combinaison : le vote de majorité, la moyenne pondérée à 50/50 et finalement une méthode d'apprentissage d'ensemble avec une couche neuronale apprenante.

Nous avons bien vérifié une de nos hypothèses, celle qui disait qu'en combinant les résultats de plusieurs systèmes gagnants de la FNC, on obtiendrait de meilleurs résultats. Par contre les résultats nous donnent une bonne leçon d'humilité en ce que nos modèles les plus complexes n'ont pas eu les meilleurs résultats. En effet, le vote de majorité ressort grand gagnant de nos améliorations alors qu'il est le système le plus simple et naïf. *A posteriori* nous expliquons son succès du fait que c'est le modèle qui dispose de plus de ressources par rapport aux autres. En effet, nos autres modèles de combinaison ou notre modèle d'apprentissage par ensemble ne se base que sur les résultats de deux sous-modèles gagnants, alors que le vote de majorité dispose des résultats de tous les participants gagnants.

Ainsi, le modèle de vote de majorité où le système Athene est dominant dépasse notre baseline de 1.26 avec le score de 83,28 au pourcentage FNC. Nous regrettons, en revanche, la difficulté d'accessibilité du code du système d'Athene pour le profane. Pour cette raison, nous n'avons pas pu produire des probabilités de labels que nos autres modèles utilisent. Ceci est l'explication de l'absence de modèles utilisant soit notre moyenne soit notre apprentissage par ensemble.

Nous regrettons aussi de ne pas avoir réussi à utiliser correctement le module TF-IDF du modèle Solat in the Swen. Cela a créé plus de variables indépendantes pour nos modèles de moyenne et notre modèle d'apprentissage d'ensemble. C'est donc moins pratique pour la comparaison de modèle.

Notre autre hypothèse : « mieux distinguer les labels agree et discuss améliore les résultats » fut vérifiée de manière indirecte. En effet, nous regrettons aussi de ne pas avoir utilisé plus de technologie du traitement du langage afin de mieux faire cette distinction. Mais à vrai dire, nous n'étions pas sûr de la voie à emprunter. Les systèmes gagnants ont exploré de manière assez exhaustive les technologies pertinentes. Vouloir proposer des traits non-significatifs n'était pas notre but. C'est pourquoi nous nous sommes concentrés sur des combinaisons de systèmes. Mais si par le futur, il venait à y avoir une découverte de traits plus significative qu'actuellement, ces systèmes ne demanderaient qu'à être améliorés.

Nous nous sommes aussi rendus compte qu'entraîner les systèmes sur les *headlines* de test ne changeait pas beaucoup l'espace sémantique des classificateurs. Les résultats sont assez proches, les modèles étant entraînés avec ou sans les *headlines*.

Chapitre 4

Conclusions

Dans ce mémoire, nous avons vu ce qu'était donc une *Fake News*, ce phénomène nouveau d'informations volontairement trompeuses qui pullulent et se reproduisent sur le web. Nous avons décrit leur origine, leur mode de fonctionnement et leur utilisation. Entre autres, nous avons vu à quel point elles peuvent être dangereuses, lorsqu'elles sont utilisées à des fins anti-démocratiques. Nous avons donc proposé une méthode manuelle pour s'en protéger de manière individuelle et éducative. Ensuite, nous sommes allés voir les systèmes informatiques préliminaires de détection de *Fake News*. Labelliser comme vrai/faux est pour le moment impossible, du fait du manque cruel de corpus d'entraînement pour ce type de tâches. Nous nous sommes donc rabattus sur les technologies de traitement du langage qui pourraient potentiellement aider à créer un classificateur *Fake News*/*True News* dans le futur. Ainsi, nous avons présenté la détection de parti pris. Nous nous sommes familiarisés avec le concept en explorant la tâche 6 semeval 2016. Puis nous nous sommes rapprochés de notre but en expliquant la *Fake News Challenge*. Nous avons discuté les différents résultats obtenus des participants gagnants au FNC et nous avons décrit leurs modèles. Dans notre dernière partie nous avons proposé nos propres modèles combinatoires pour répondre à la *Fake News Challenge*. De manière minime, nous avons dépassé l'état de l'art.

En conclusion, nous avons donc atteint nos trois objectifs, à savoir définir exhaustivement ce qu'est une *Fake News*, expliquer comment la détection de parti pris peut aider à leur repérage et dépasser l'état de l'art en la matière. Par contre, nous avons eu seulement une approche très quantitative et orientée par l'organisation du *Fake News Challenge*. Nous n'avons rien proposé de plus qui n'ait été partiellement exploré auparavant.

Or, pour détecter les *Fake News*, comme nous l'avons dit plus tôt, il faut s'intéresser à la notion de vérité et d'adéquation à la réalité. C'est-à-dire qu'il faudrait plutôt avoir une approche qualitative. Ici, nous avons uniquement produit un classificateur de relation entre un titre et un texte. Nous n'avons aucun indice sur la nature de l'information qui permettrait de détecter les *Fake News*. Nous savons juste si le titre est en accord, en désaccord, lié ou discutant le texte. Cela ne nous aide pas à connaître la véracité du titre ou du texte. Au mieux, si nous arrivons à déterminer la véracité d'un élément, nous pourrons déduire la véracité partielle ou complète du second. En somme, par déduction logique et avec la véracité d'une partie, nous aurions pu faire des affirmations sur la réalité. Si nous pouvions disposer d'un convertisseur fonctionnel du langage naturel vers des énoncés logiques, sans ambiguïté

sémantique, alors nous pourrions déterminer plus formellement la validité des énoncés à vérifier par rapport à des corpus dont nous connaîtrions la véracité des éléments. Or, construire un tel corpus est très chronophage. Nous pourrions certes semi-automatiser son fonctionnement avec la *Stance Detection* mais l'intervention humaine de vérification constituerait une grande partie du travail. De plus, pour que nous ayons des résultats satisfaisants et exploitables, il nous faudrait beaucoup d'exemples de *Fake News*. Et en réalité, les *Fake News* profitent de beaucoup plus de visibilité que les *True News* mais elles demeurent beaucoup moins nombreuses. Pour que le système soit le plus performant possible il lui faudrait beaucoup d'échantillons de *Fake News* pour savoir les identifier et encore mieux distinguer les *True News* de ces dernières.

Ce mémoire nous montre combien la tâche de détection de *Fake News* est difficile. Nous n'avons ici qu'un proto-détecteur de *Fake News* qui n'est pas garant de la véracité mais qui peut établir des relations entre des titres et des corps de texte de manière relativement satisfaisante.

Bibliographie

[1] Alcaraz, M. (2018). Concert annulé de Bertrand Cantat : des associations comptaient manifester dans l'Olympia.

[2] Amanda-Mary (2018). NOW IT'S OFFICIAL : FDA Announced That Vaccines Are Causing Autism !

[3] Anand, P., Walker, M., Abbott, R., Tree, J. E. F., Bowmani, R., and Minor, M. (2011). Cats rule and dogs drool ! : Classifying stance in online debate. In *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

[4] Arrigoni, G. (2017). Johnny Hallyday est mort à l'âge de 74 ans.

[5] Baird, S., Sibley, D., and Pan, Y. (2017). Talos Targets Disinformation with Fake News Challenge Victory.

[6] Barbouch, R. (2018). Accusations contre Médiapart : Edwy Plenel répond à Nicolas Sarkozy.

[7] Bienvault, P. (2018). La rougeole continue à être mortelle en France.

[8] Blanchard, O. (2017). Des anti-immigration prennent des sièges de bus pour des femmes voilées.

[9] Cometti, L. (2018). Vœux à la presse : Macron veut censurer les « Fake News» et prône une «saine distance» entre pouvoir et médias.

[10] de Ficientis, F. C. (2018). Un politicien brésilien corrompu accroché à un poteau !

[11] Ferreira, W. and Vlachos, A. (2016). Emergent : a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.

[12] Freeman, L. (2017). Le Président Duterte banni les vaccins aux Philippines : « Les Vaccinations provoquent l'Autisme ».

[13] Gabielkov, M., Ramachandran, A., Chaintreau, A., and Legout, A. (2016). Social clicks : What and who gets read on twitter ?

[14] Gauron, R. (2017). «Fake News», un même terme pour plusieurs réalités.

[15] Hanselowsk, A., PVS, A., Schiller, B., and Caspelherr, F. (2017). Team Athene on the Fake News Challenge.

[16] Harcup, T. and O'Neill, D. (2016). What is News ? News values revisited (again).

[17] Jacobson, L. (2016). Yes, Donald Trump did call climate change a Chinese hoax.

[18] Maclin, R. and Opitz, D. W. (2011). Popular ensemble methods : An empirical study. *CoRR*, abs/1106.0257.

[19] Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). Semeval-2016 task 6 : Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

[20] Parlanti, C. (2014). Christine Boutin cite le Gorafi sur BFMTV : si j'étais pro Manif pour tous, j'aurais honte.

[21] Reichstadt, R. (2018). Le conspirationnisme dans l'opinion publique française.

[22] Riedel, B., Augenstein, I., Spithourakis, G. P., and Riedel, S. (2017). A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *CoRR*, abs/1707.03264.

[23] [s. a.] (2014). Barilla et Doliprane lancent des pâtes spéciales « fin de soirée » avec des vrais morceaux d'aspirine.

[24] [s. a.] (2016). Erratum.

[25] [s. a.] (2017a). Alain Finkielkraut : « À ma mort, je léguerai tout mon patrimoine aux migrants ».

[26] [s. a.] (2017b). Donald Trump sort les États-Unis de l'Accord de Paris sur le climat.

[27] [s. a.] (2017c). Les cas de rougeole ont doublé en une année.

[28] [s. a.] (2017d). L'usine à Fake News : comment des écoliers macédoniens ont «élu» Trump président des USA.

[29] [s. a.] (2017e). Ouragan Maria. L'état de catastrophe naturelle élargi à 15 communes de Guadeloupe.

[30] [s. a.] (2017f). Pourquoi tout le monde - ou presque - partage des fake news ?

[31] [s. a.] (2017g). Unos 5 diputados heridos dejó incursión de oficialistas en Parlamento venezolano.

[32] [s. a.] (2017h). Vaccination et autisme : « Nous réclamons justice et réparation pour nos enfants blessés par leur vaccin ».

[33] [s. a.] (2018a). Alexia Mori enceinte : Elle dévoile combien de kilos elle a pris !

[34] [s. a.] (2018b). Héritage de Stephen Hawking : Sa famille se dispute pour savoir qui obtiendra le trou noir dans son garage.

[35] Sherman, A. (2017). No, the FDA didn't hide information linking vaccine to autism.

[36] Sobhani, P., Mohammad, S., and Kiritchenko, S. (2016). Detecting stance in tweets and analyzing its interaction with sentiment. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 159–169, Berlin, Germany. Association for Computational Linguistics.

[37] Wei, W., Zhang, X., Liu, X., Chen, W., and Wang, T. (2016). pkudblab at semeval-2016 task 6 : A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 384–388, San Diego, California. Association for Computational Linguistics.

[38] Zannettou, S., Caulfield, T., Cristofaro, E. D., Kourtellis, N., Leontiadis, I., Sirivianos, M., Stringhini, G., and Blackburn, J. (2017). The web centipede : Understanding how web communities influence each other through the lens of mainstream and alternative news sources. *CoRR*, abs/1705.06947.

[39] Zarrella, G. and Marsh, A. (2016). Mitre at semeval-2016 task 6 : Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 458–463, San Diego, California. Association for Computational Linguistics.

Annexe A

Les soumissions complètes des gagnants du FNC

Stance	unrelated	agree	disagree	discuss	Somme
Nombre	18349	1903	697	4464	25413
Pourcentage	72.20%	7.48%	2.74%	17.56%	100%
Score FNC	4587.25	1903	697	4464	11651.25

TABLE A.1 Répartition des données dans l' ensembe de test.

Solat					
	agree	disagree	discuss	unrelated	Somme
agree	1114	17	588	184	1903
disagree	275	13	294	115	697
discuss	823	6	3401	234	4464
unrelated	35	0	203	18111	18349
Somme	2247	36	4486	18644	25413

TABLE A.2 Table de confusion du modèle : Solat

Solat				
Mesure	agree	disagree	discuss	unrelated
Précision	0.59	0.02	0.76	0.99
Rappel	0.5	0.36	0.76	0.97
F1score	0.54	0.04	0.76	0.98
Exactitude	89.08			
Score FNC	9556.5			
Pourcentage FNC	82.02			

TABLE A.3 Mesures pour le modèle : Solat

Athene

	agree	disagree	discuss	unrelated	Somme
agree	851	68	827	157	1903
disagree	241	66	241	149	697
discuss	466	37	3611	350	4464
unrelated	19	4	115	18211	18349
Somme	1577	175	4794	18867	25413

TABLE A.4 Table de confusion du modèle : Athene

Athene

Mesure	agree	disagree	discuss	unrelated
Précision	0.45	0.09	0.81	0.99
Rappel	0.54	0.38	0.75	0.97
F1score	0.49	0.15	0.78	0.98
Exactitude	89.48			
Score FNC	9550.75			
Pourcentage FNC	81.97			

TABLE A.5 Mesures pour le modèle : Athene

UCL-mr

	agree	disagree	discuss	unrelated	Somme
agree	838	12	939	114	1903
disagree	179	46	356	116	697
discuss	523	46	3633	262	4464
unrelated	53	3	330	17963	18349
Somme	1593	107	5258	18455	25413

TABLE A.6 Mesures pour le modèle : UCL-mr

UCL-mr

Mesure	agree	disagree	discuss	unrelated
Précision	0.44	0.07	0.81	0.98
Rappel	0.53	0.43	0.69	0.97
F1score	0.48	0.11	0.75	0.98
Exactitude	88.46			
Score FNC	9521.5			
Pourcentage FNC	81.72			

TABLE A.7 Mesures pour le modèle : UCL-mr