ANALYSE DE DONNÉES TEXTUELLES EN PSYCHIATRIE

TEXT ANALYSIS APPLICATIONS TO PSYCHIATRY

Louis FALISSARD*

RÉSUMÉ

Avec la démocratisation des réseaux sociaux et des applications de messagerie, une grande partie de nos interactions sociales se sont digitalisées, notamment vers un format textuel. L'analyse quantitative de ces jeux de données textuelles massifs, grâce à des techniques de machine learning, comme le deep learning notamment, a engendré de nombreuses innovations technologiques qui sont de nos jours quotidiennement utilisées par des millions d'utilisateurs. Le diagnostic en psychiatrie étant basé, au moins en partie, sur le langage, il semble naturel de s'intéresser à l'utilisation de ces techniques d'analyse de données textuelles dans une approche de santé mentale. Il est donc nécessaire de s'interroger sur les différentes approches méthodologiques et analytiques disponibles, ainsi que leurs potentielles applications dans le domaine de la psychiatrie.

MOTS-CLÉS

Traitement du language naturel, Psychiatrie, Analyse de données, Apprentissage machine, Apprentissage profond, Réseaux de neurones artificiels.

ABSTRACT

The advent of social network and messenger apps has led to the digitalization of a significant part of our social interactions, specifically toward a textual format. The quantitative analysis of these huge textual dataset, through the use of machine learning, and more specifically deep neural networks, led to numerous technological innovations which are now used daily by millions of users. As diagnosis in psychiatry is based, at least partly, on language, it only seems natural to investigate the use of these technologies in mental health. Consequently, an interrogation on the different available approaches, whether methodological or analytical, as well as their potential applications in psychiatry, is necessary.

KEYWORDS

Natural language processing, Psychiatry, Data analysis, Machine learning, Deep learning, Artificial neural networks.

* * *

INTRODUCTION

Durant ces dernières années, le domaine de l'apprentissage machine (machine learning) a connu une expansion considérable conduisant à de multiples applications pratiques maintenant utilisées par des millions d'utilisateurs, ce dans des domaines aussi variés que l'analyse d'image, de voix, ou encore de langage. Ces récentes innovations sont dues principalement aux développements d'une sous-discipline du machine learning, l'apprentissage profond (deep learning), et à sa capacité d'analyser des jeux de données complexes et par le biais de représentations internes quantitatives et hiérarchique de complexité et puissance explicative croissante. Plus

* Oxford University, Nuffield Department of Surgical Sciences, Computational Neuroscience lab - louis.falissard@gmail.com

précisément, les architectures de réseaux de neurones artificiels proposées en apprentissage profond constituent un environnement de modélisation mathématiques qui s'adapte naturellement à la modélisation de relations non-linéaires présentes dans un jeu de données, tout autant qu'à l'analyse de données structurées comme les données longitudinales ou séquentielles, notamment par le biais de l'utilisation de réseaux de neurones récurrents. Ces derniers ont été notamment à la source de progrès considérables en analyse de texte, et ceci dans des contextes aussi variés que la classification de document à la traduction automatique, en passant par la construction d'intelligences artificielles spécialisée dans l'intervention en chat.

D'un autre côté, les interactions sociales dans les populations occidentales ont connu au cours de ces dernières années une migration considérable vers des formats digitalisés, sous un format majoritairement textuel. En effet, les réseaux sociaux présentent à leurs membres un médium d'expression et de partage de leurs émotions et expériences à un niveau jamais observé auparavant. Les applications de messageries offrent à leurs utilisateurs des possibilités constantes et continues de communications avec leur entourage. Toute cette expression de la personnalité et l'état d'un individu s'effectuant alors par le biais de publications et d'échanges textuels, dont le niveau de privatisation varie en fonction des technologies utilisées.

Cet article s'intéresse à l'application de ces puissantes méthodes d'analyse sur les données textuelles fournies par ces nouvelles technologies, ceci dans une approche de santé mentale. Si l'utilisation de méthodes quantitatives en psychiatrie ne se fait pas naturellement, il est possible par le biais de méthodologies bien construites, de construire des jeux de données permettant de possibles innovations dans la prise en charge des troubles mentaux par le clinicien. La première partie définie une méthodologie précise de construction d'un jeu de données adapté au problème de la psychiatrie, étape cruciale dans l'implémentation d'un algorithme de machine learning. La seconde partie présente une méthode d'utilisation des techniques d'apprentissage profond pour l'analyse de données textuelles, et notamment leur utilisation en analyse de régression.

MÉTHODOLOGIE

L'analyse de texte est une discipline très riche, et de multiples approches à ce problème ont été développées au fil des années. Cet article se concentre sur l'analyse de texte basée sur l'apprentissage machine. La plupart des implémentations d'algorithmes de machine learning nécessite la construction préalable d'un jeu de données, semblable de nature aux jeux de données collectés en vue d'une analyse de régression. Deux concepts ont donc besoin d'être définis avant de procéder à la récolte du jeu de données :

- Les variables explicatives (typiquement, le type de texte à étudier, sa provenance, etc.).
- La variable à expliquer (typiquement une variable quantitative ou catégorielle, dans ce cadre d'application rendant compte de l'état émotionnel ou mental du patient) associée aux variables explicatives.

L'élaboration de ces variables a un impact considérable sur la qualité de l'algorithme implémenté, ainsi que sur le coût potentiel de l'étude. En effet, l'apprentissage profond requiert en règle générale des tailles d'échantillons considérables (>5 000).

I. VARIABLES EXPLICATIVES

Type de données

Bien qu'omniprésente en ligne, les données textuelles ne sont pas nécessairement faciles à extraire, et l'information extraite peut potentiellement s'avérer plus riche que le texte seul. Bien que particulièrement riches, les données d'utilisation d'une messagerie instantanée, par exemple, sont, en raison des mécanismes de cryptage qu'elles incorporent, difficile à collecter. L'utilisateur est cependant facilement capable de les extraire sous un format adapté à l'analyse. La figure 1 présente le format de données textuelles typiquement obtenu à partir de ces messageries.

En plus de contenir les données textuelles des conversations de l'individu, ce format présente d'autres informations qui peuvent s'avérer particulièrement utiles en analyse, les métadonnées. En effet, l'heure d'envoi des messages, par exemple, est une information plus facilement quantifiable ; et donc analysable, que le texte en lui-même, et peut potentiellement apporter une information précieuse (fréquence d'envoi de messages très élevée à un instant donné, heure d'envoi inhabituelle pour une personne donnée, etc.).

Environnement des données

Le choix de l'environnement de collecte des données joue un rôle considérable dans l'élaboration du jeu de données, et a un impact sur plusieurs facteurs. Typiquement, deux principaux modes de collecte peuvent être identifiés :

- Texte extrait dans un environnement libre (application de messagerie, réseaux sociaux, etc.).

```
04/02/2017, 21:56 - A: [...]
04/02/2017, 21:56 - A: [...]
04/02/2017, 22:17 - B: [...]
04/02/2017, 22:17 - B: [...]
04/02/2017, 22:17 - B: [...]
04/02/2017, 23:25 - A: [...]
04/02/2017, 23:25 - A: [...]
04/02/2017, 23:26 - A: [...]
05/02/2017, 07:29 - B: [...]
```

Figure 1. Exemple de données textuelles obtenues sur l'application de messagerie Whatsapp.

– Ce mode d'acquisition nécessite typiquement des tailles d'échantillons plus larges en raison de la considérable variabilité inter-sujet potentiellement.
– Les données sont obtenues à partir d'application très largement utilisées, et dispose de ce fait d'une accessibilité massive.
– Texte issu d'une situation contrôlée (ex. réponse à des questions prédéfinies).
– La structure du contenu est cadrée, ce qui permet de potentiellement diminuer la variabilité inter-sujet et d'avoir directement accès à une information impactant le critère à expliquer défini, en fonction de la définition du cadre.
– L'acquisition des données est coûteuse. Il est nécessaire de recueillir manuellement chaque observation auprès d'un individu.

Variable à expliquer

C'est le choix de la variable à expliquer qui définit la nature de l'information que l'algorithme d'apprentissage profond tentera de modéliser. En effet, la fonction de cet indicateur est de donner une information quant à l'état mental du patient, que l'algorithme d'apprentissage essaiera par la suite d'approximer. Sa définition a donc un impact crucial, puisqu'il définira à lui seul le sens et l'information qui sera étudiée au cours de l'étude. Il est donc nécessaire de l'élaborer avec soin. Le choix de ce critère, autant dans sa modalité que dans son application, n'a pratiquement aucune limite. Il est cependant préférable de se cantonner à certaines situations, potentiellement plus simples d'accès et plus abordable pour le processus de modélisation.
La première étape d'élaboration de la variable à expliquer consiste à choisir le type de variable utilisé pour représenter l'état mental de l'individu :

- Variable qualitative (typiquement une décision binaire, « le sujet étudié est dépressif/non dépressif »).
- Variable quantitative (typiquement une échelle d'évaluation ordinale, par exemple un critère CGI-S [1]).
- Variable structurée (ex. questionnaire).

La définition du type de critère doit s'effectuer conjointement avec le type d'évaluation correspondant :

- Auto-évaluation.
- Ne nécessite pas l'intervention d'un praticien, donc plus économique.
- L'évaluation n'est alors pas contrôlée par un clinicien, ce qui peut être source de bruit.
- Évaluation par un clinicien.
- L'information à modéliser profite d'une expertise qui lui amène potentiellement une qualité supérieure.
- Le processus d'évaluation nécessite alors l'évaluation du clinicien sur tous les exemples recueillis, se comptant en général dans les milliers à dizaine de milliers. Il s'agit d'un processus long et rébarbatif.

Une fois la méthodologie définie, peut commencer la collecte des données. Il est difficile de déterminer a priori la taille d'échantillon nécessaire au bon fonctionnement de l'algorithme d'apprentissage profond. Il est en revanche déconseillé d'appliquer ces modèles mathématiques sur des jeux de données comptant moins de 5 000 individus.

II. APPRENTISSAGE PROFOND ET ANALYSE DE DONNÉS TEXTUELLES

Régression linéaire

L'analyse de régression linéaire est un outil de modélisation largement utilisé qui consiste à représenter un jeu de donnée par une approximation multivariée linéaire du phénomène considéré. Cette approche simple, utilisée autant en statistiques qu'en data science, partage certaines similarités avec les réseaux de neurones qui en font une bonne introduction au sujet.

L'objectif d'une analyse de régression linéaire est d'ajuster une approximation linéaire paramétrique en W à un groupe de N observations $\{(X_i, y_i)\}_{0<i<N+1}$ telle que :

$$\hat{y} = W_e \cdot X + b$$

$$with \quad W_e = \underset{W}{\operatorname{argmin}}\left(\frac{1}{N} \cdot \sum_{i=1}^{N} (\hat{y}_i - y_i)\right)$$

L'obtention de la solution paramétrique optimale We peut se faire de manière analytique, ou bien encore par le biais d'un algorithme d'optimisation comme une descente de gradient. Les modèles linéaires ainsi obtenus sont connus pour être simple à ajuster, ainsi que pour leur importante capacité de généralisation. Ils sont en revanche incapable d'exprimer des relations non-linéaires ainsi que des interactions inter-variables.

Perceptron multi-couches

Les perceptrons multicouches représentent une des familles de réseaux de neurones artificiels les plus utilisée parmi les modèles de deep learning [2], et peuvent être utilisés pour une variété de tâches, notamment en analyse de régression, où ils peuvent être interprétés comme une extension non-linéaire à l'analyse en régression linéaire susmentionnée.

L'idée générale derrière un perceptron multicouche consiste, de manière similaire au processus d'analyse de régression linéaire, à ajuster un modèle linéaire. Cependant, celui-ci sera ajusté à une transformée non linéaire du groupe d'observation $\{(\Phi(X_i), y_i)\}_{0<i<N+1}$. La puissance d'un perceptron multicouches, notamment par rapport aux méthodes d'expansions de base couramment utilisée en modélisation statistique, réside dans le fait qu'en plus de déterminer les paramètres du modèle linéaire à partir des observations, la transformée Φ l'est également.

Pour permettre à un perceptron multicouche d'apprendre des interactions non-linéaire, l'approche traditionnellement utilisée consiste à injecter des combinaisons linéaires des variables étudiées dans de simples non-linéarités, les résultats obtenus étant alors utilisés soit comme variables explicatives pour un modèle linéaire, soit combinées et réinjectées dans de nouvelles linéarités. Le réseau simplifié présenté en figure 1 est un exemple de perceptron multicouche adapté à l'analyse de régression. De la même manière que pour une analyse de régression linéaire, l'objectif est d'ajuster une approximation paramétrique, cette fois ci non-linéaire, à un groupe de N observations $((x_1, x_2)_i, y_i)_{0<i<N+1}$ tel que :

$$\hat{y} = w_{11}^3 \cdot f^1(w_{11}^2 \cdot x_1 + w_{12}^2 \cdot x_2 + b_1^2) + w_{12}^3 \cdot f^2(w_{21}^2 \cdot x_1 + w_{22}^2 \cdot x_2 + b_2^2) + b_1^3$$

Si les paramètres sont ici typiquement déterminés par le biais de la méthode des moindres carrés, deux différences notables découlent de l'introduction des non-linéarités paramétriques dans le modèle :

- La solution paramétrique optimale ne peut plus s'exprimer sous forme fermée. Il est nécessaire de l'estimer par le biais d'un algorithme du gradient.
- L'obtention du gradient de la fonction objectif ne peut plus se faire directement, mais reste possible, notamment par l'utilisation de techniques de rétro-propagation du gradient [3].
- Le problème d'optimisation ainsi défini perd la propriété de convexité présente en analyse de régression linéaire, et avec elle la garantie de parvenir à déterminer le minimum global de la fonction objectif [4].

Tout perceptron multicouche peut ainsi s'obtenir de cet exemple simplifié, principalement par la variation de trois facteurs :

- Le nombre de fonctions composées enchainées (appelé nombre de couches de neurones).
- Le nombre de variable artificielles créées dans chaque couches (nombre de neurones par couches).
- Le type de non-linéarités utilisée à la sortie de chaque neurone (i.e. le choix des fonctions f1 et f2).

Des modèles de complexités croissante peuvent ainsi être construits en adaptant ces trois facteurs, permettant ainsi la construction d'un modèle à la puissance explicative adaptée à la complexité du problème.

Si en théorie tout type de fonction non-linéaire peut être utilisée à la sortie d'un neurone, il est vivement recommandé de limiter le choix de ses non-linéarités à certaines options dont le bon comportement a été, si non théoriquement démontré, au moins vérifié expérimentalement, comme par exemple la fonction tangente hyperbolique et la fonction linéaire rectifiée.

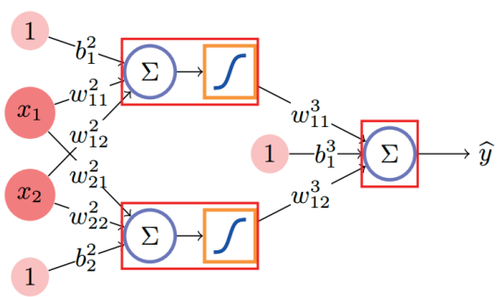

Figure 2. Exemple de perceptron multicouche adapté à l'analyse regressive.

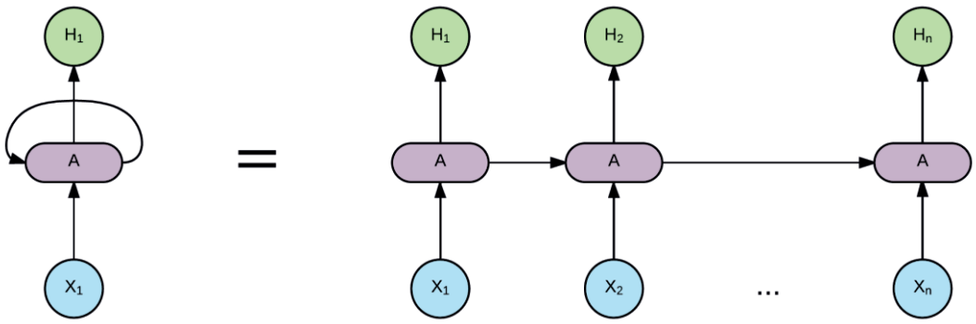

Figure 3. Réseaux de neurones récurrent, avec sur la droite sa forme avec boucle, et sur la gauche sa représentation déroulée.

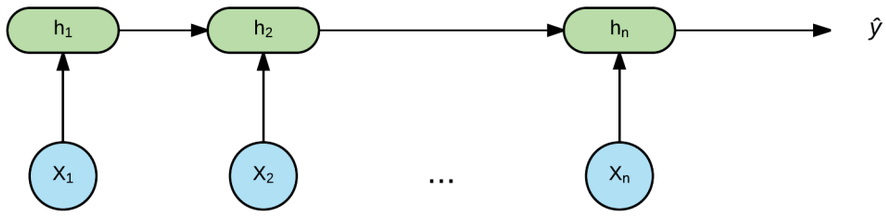

Figure 4. Réseaux de neurones recurrent adapté a l'analyse de regression.

Analyse de séquences et Réseau de Neurones récurrent

Un perceptron multicouche permet théoriquement l'analyse de séquences de données $x^{(1)}$, $x^{(2)}$,…,$x^{(n)}$, par exemple en considérant la séquence comme une observation multivariée et en l'injectant directement dans le réseau de neurones. Cette approche présente cependant des problèmes. En effet, le perceptron tel qu'il est traditionnellement défini voit son nombre de paramètre augmenter drastiquement avec la taille de la séquence étudiée. De plus, un tel modèle ne permet clairement pas l'analyse de séquences de longueur variable, ce qui se montre prohibitif dans un nombre de situations dont l'analyse de texte fait partie.

Les réseaux de neurones récurrents constituent une famille de réseaux de neurones artificiels spécialisés dans l'analyse de séquences de données [2]. L'idée sous-jacente aux réseaux de neurones récurrent consiste à élaborer un modèle qui partage ses paramètres à travers l'évolution séquentielle des variables observées. Ainsi, au lieu d'injecter directement la séquence entière au réseau, chacun de ses éléments est présenté séquentiellement au réseau, qui, pour tenir compte des observations passées, est doté d'une boucle de rétroaction.

Tout comme les perceptrons multicouches, les réseaux de neurones récurrents peuvent être utilisés de multiple manière, comme par exemple en analyse régressive de données séquentielles. L'objectif est alors d'ajuster une approximation paramétrique à un groupe de N observations de séquences $((X_1, ..., X_n)_i, y_i)_{1<i<N}$ telle que :

$$\forall i \in \mathbb{N}, i \in [1, n]$$
$$h_i = f(x_i, h_{i-1}) \quad \textit{avec } f \textit{ un membre d'une famille de fonctions paramétriques}$$
$$\hat{y} = \mathrm{W} \, . \, h_n$$

De manière similaire au perceptron multicouche, ŷ peut être approximé par une analyse de régression linéaire effectuée sur la dernière sortie du réseau de neurones récurrent. Les paramètres de cette dernière, ainsi que de la fonction f, historiquement définie comme un perceptron multicouche, étant alors déterminés par méthode des moindres carrés.

Pour des raisons qui dépassent le cadre de cette article, l'utilisation d'un perceptron multicouche comme définition de la fonction paramétrique f présente cependant plusieurs défauts qui résultent en des difficultés considérables au cours du processus d'optimisation [4].

Plusieurs architectures ont depuis été développées pour remédier à ces problèmes, tout comme les *Gated Recurrent Units*, ou encore les cellules *Long-Short-Term-Memory* (LSTM), cette dernière étant celle retenue dans cet article. La cellule LSTM [5] est une variation dans l'utilisation du perceptron multicouche comme choix de la fonction paramétrique f dans la construction d'un réseau de neurones récurrents, celle-ci étant alors définie telle que :

$$f_t = \sigma_g(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f)$$

$$i_t = \sigma_g(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i)$$

$$o_t = \sigma_g(W_o \cdot x_t + U_o \cdot h_{t-1} + b_o)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_g(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f)$$

$$h_t = o_t \circ \sigma_h(c_t)$$

Avec :

- x_t le vecteur d'entrée (variables observées)
- h_t le vecteur de sortie
- c_t le vecteur d'état de cellule
- W, U, et b les paramètres de la cellule
- σg la fonction logistique
- σc et σh la fonction tangente hyperbolique

APPLICATION DES RÉSEAUX DE NEURONES RÉCURRENT À L'ANALYSE DE RÉGRESSION SUR DONNÉES TEXTUELLES

Considérons un corpus de documents (X_1, ..., X_N), chacun associé à une variable à expliquer y. Comme précédemment abordé, y correspond à un critère prédéfini au cours de la construction du jeu de données, et peut par exemple représenter l'impression d'un praticien quant à la présence d'élément de pulsions suicidaires par le biais d'une évaluation CGI-S. Chaque document est constitué d'une séquence de mots Xi = (X_1, ..., X_T) avec T la taille du document Xi. L'objectif est donc d'ajuster une approximation paramétrique à un groupe de N observations de documents $((X_1, ..., X_n)_i, y_i)_{1<i<N}$, correspondant typiquement aux conditions précédemment définies d'analyse régressive de données séquentielles.

Quand bien même les réseaux de neurones récurrents proposent un environnement de modélisation qui s'adaptent naturellement à l'analyse de séquences, leur application à l'analyse de données textuelles nécessite un travail supplémentaire de conversion. En effet, bien qu'un texte puisse se considérer comme une séquence de mots, ces derniers ne constituent en aucun cas des variables quantitatives. Lors d'un exercice de classification ou d'analyse de régression, il existe une manière simple de transformer un jeu de données textuel en jeu de données quantitatives longitudinales [6] :

Un vocabulaire V constitué d'un nombre de mots fixé est défini.

Un dictionnaire D associant chaque mot à un vecteur quantitatif (dimensionnalité typiquement comprise entre 50 et 500) initialisé aléatoirement est défini.

Les textes du jeu de données sont convertis en séquences de variables quantitatives en utilisant le dictionnaire susmentionné.

Un réseau de neurone récurrent est ajusté sur la tâche d'analyse de régression prédéfinie, comme décrit précédemment.

L'optimisation du modèle se fait à la fois sur les paramètres du réseau de neurones, et sur le dictionnaire D utilisé pour convertir les données textuelles.

Intuitivement, l'optimisation conjointe du réseau de neurone et du dictionnaire texte/quantitatif permet au réseau de neurones de transformer le dictionnaire D, initialisé aléatoirement, en une représentation quantitative du vocabulaire V encapsulant l'information nécessaire à l'ajustement des paramètres du réseau de neurone récurrent destiné à l'analyse de régression.

CONCLUSION

Les récentes avancées dans l'utilisation des réseaux de neurones dans les processus de modélisation de données hiérarchique ont été à l'origine d'une véritable révolution dans le domaine de l'analyse de données textuelles, que ce soit dans le domaine de la classification de document, la traduction automatique ou encore l'analyse exploratoire. Leur application dans un cadre de santé mentale reste complexe. En effet, les notions mises en jeu dans la compréhension du discours d'un sujet pensant en visée d'établir un diagnostic de trouble mental sont, à l'heure qui l'est, trop complexes pour être appréhendées par ces outils mathématiques. Cependant, en les restreignant à des contextes spécifiques, et en élaborant des méthodologies adaptées, l'application des réseaux de neurones à la psychiatrie, notamment en analyse de texte, peut potentiellement donner naissance à tout une famille d'outil quantitatifs capable d'accompagner le clinicien dans son travail de diagnostic, ses prises de décisions et son suivi du patient. En raison du caractère récent de la naissance de cette nouvelle discipline d'analyse quantitative, très peu de travaux ont été réalisés dans l'optique de caractériser leur performance dans le cadre d'une application en santé mentale. Il est donc maintenant nécessaire pour praticiens et data-scientists de commencer ensemble à explorer ces notions, et à amorcer une réflexion sur la place potentielle de ces outils technologiques en santé mentale. ■

BIBLIOGRAPHIE

[1] J. Busner and S. D. Targum, "The Clinical Global Impressions Scale," *Psychiatry Edgmont*, vol. 4, no. 7, pp. 28–37, Jul. 2007.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[3] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[4] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[5] "Long Short-Term Memory | Neural Computation | MIT Press Journals." [Online]. Available: http://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.8.1735. [Accessed: 07-Sep-2017].

[6] D. Yogatama, C. Dyer, W. Ling, and P. Blunsom, "Generative and Discriminative Text Classification with Recurrent Neural Networks," *ArXiv170301898 Cs Stat*, Mar. 2017.