ANNEXE

Tableau 1 Pourcentages de variance expliquée des résultats d'une SVD réduisant à un nombre de dimensions données

Nombre de dimensions souhaitées, *n*	Pourcentage de variance expliquée (%)
100	9,89
500	27,35
1 000	39,64
10 000	82

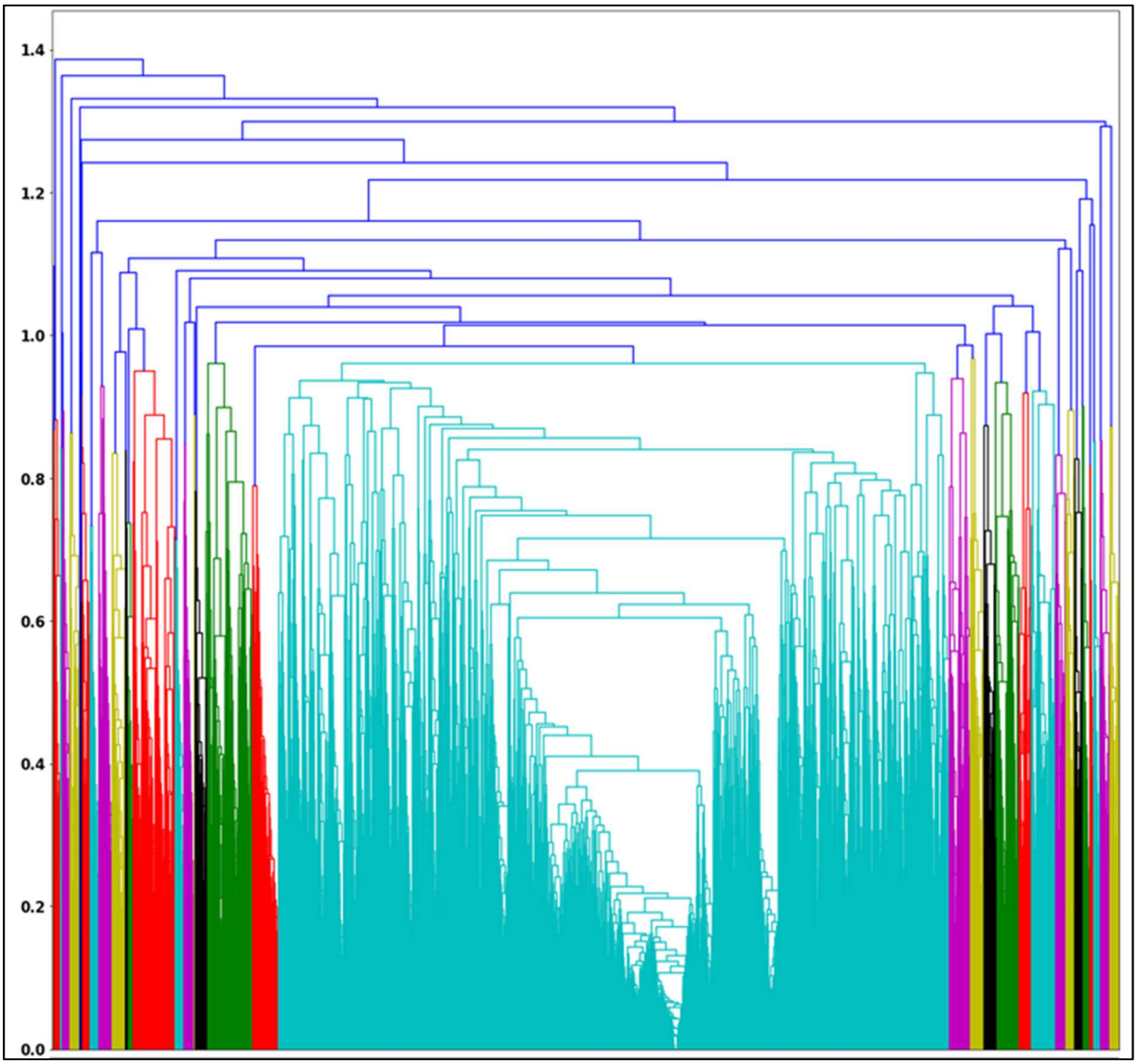

Figure 1 Dendrogramme illustrant le résultat d'un *complete-linkage* sur 10 000 éléments de la matrice TF-IDF. (Réduction à 100 dimensions par SVD, utilisation du calcul des distances euclidiennes)

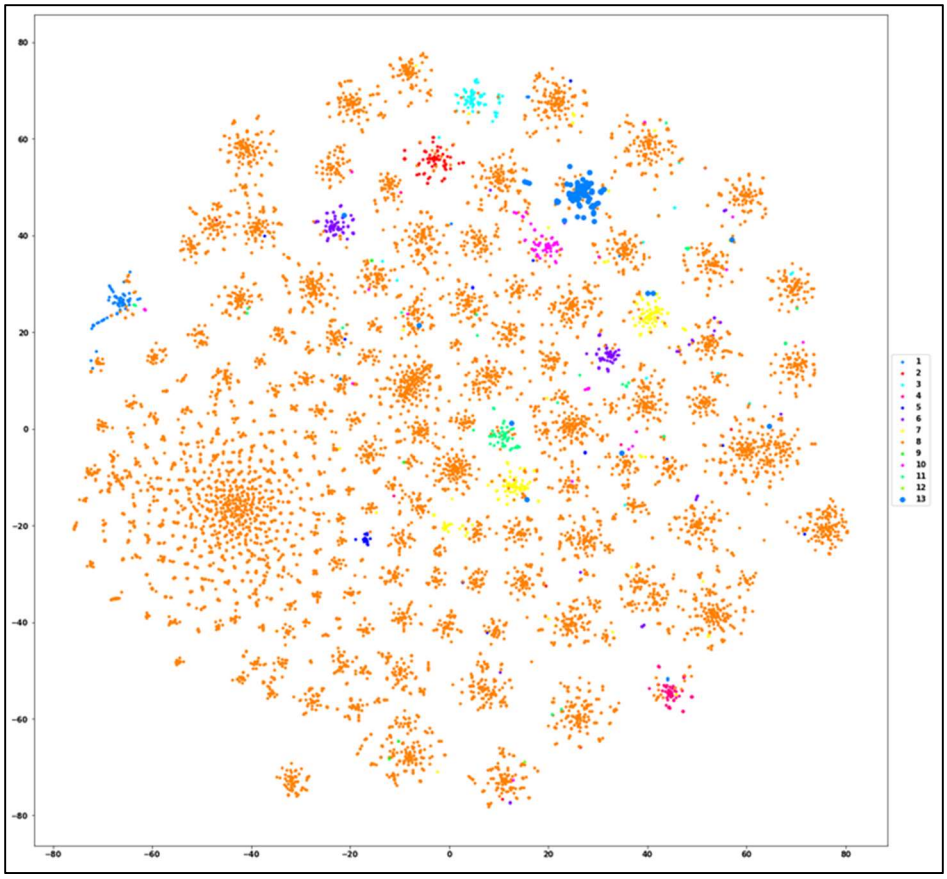

Figure 2 t-SNE affichant en couleurs 13 clusters d'un *complete-linkage* (une SVD à 100 dimensions, un calcul de distance euclidienne et les hyperparamètres *perplexity*=34 et *early_exageration*=3 ont été utilisés pour t-SNE)

Tableau 2 Identification des thèmes de certains amas et illustration de la méthodologie employée

Expérience (K)	Numéro d'amas	Nombre de tweets	Thème	IDF	Jeton	% de tweets dans l'amas contenant le jeton
50	48	571	Corée du Nord	5,63	north	0,88
				6,30	korea	0,71
				5,85	south	0,17
350	105	808	Feux de forêt, renvois d'employés	5,28	fire	1,00
350	125	369	Nouvelles mondiales, coupe du monde de football	5,49	world	1,00
				7,10	cup	0,20
350	22	32	Plan de soins de santé proposé par les républicains	5,20	plan	1,00
				5,88	health	0,88
				6,30	care	0,38
				6,27	gop	0,22
350	30	41	Suggestions d'activités à faire lors de la fin de semaine	5,86	weekend	1,00
				5,79	thing	0,54
				6,82	fun	0,12

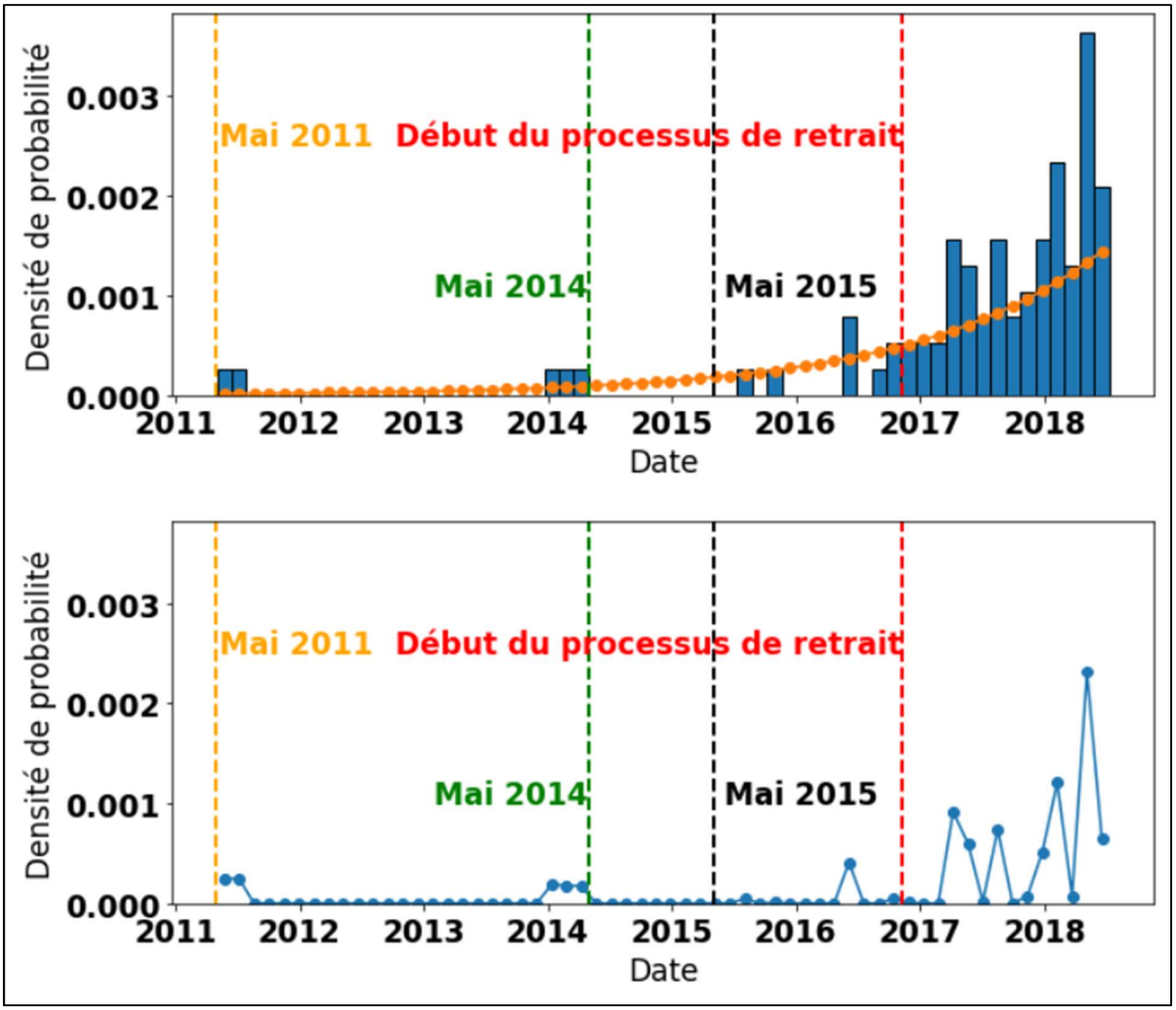

Figure 3 Distribution des tweets de l'amas dont 100% des tweets contiennent le mot **may**, 56% contiennent le mot **brexit** et 7% contiennent le mot **theresa**

Le graphique du haut présente la distribution brute ainsi que la tendance exponentielle (orange) de l'ensemble de la base de données. La figure du bas présente la distribution après le retrait de la tendance de l'ensemble de la base de données.

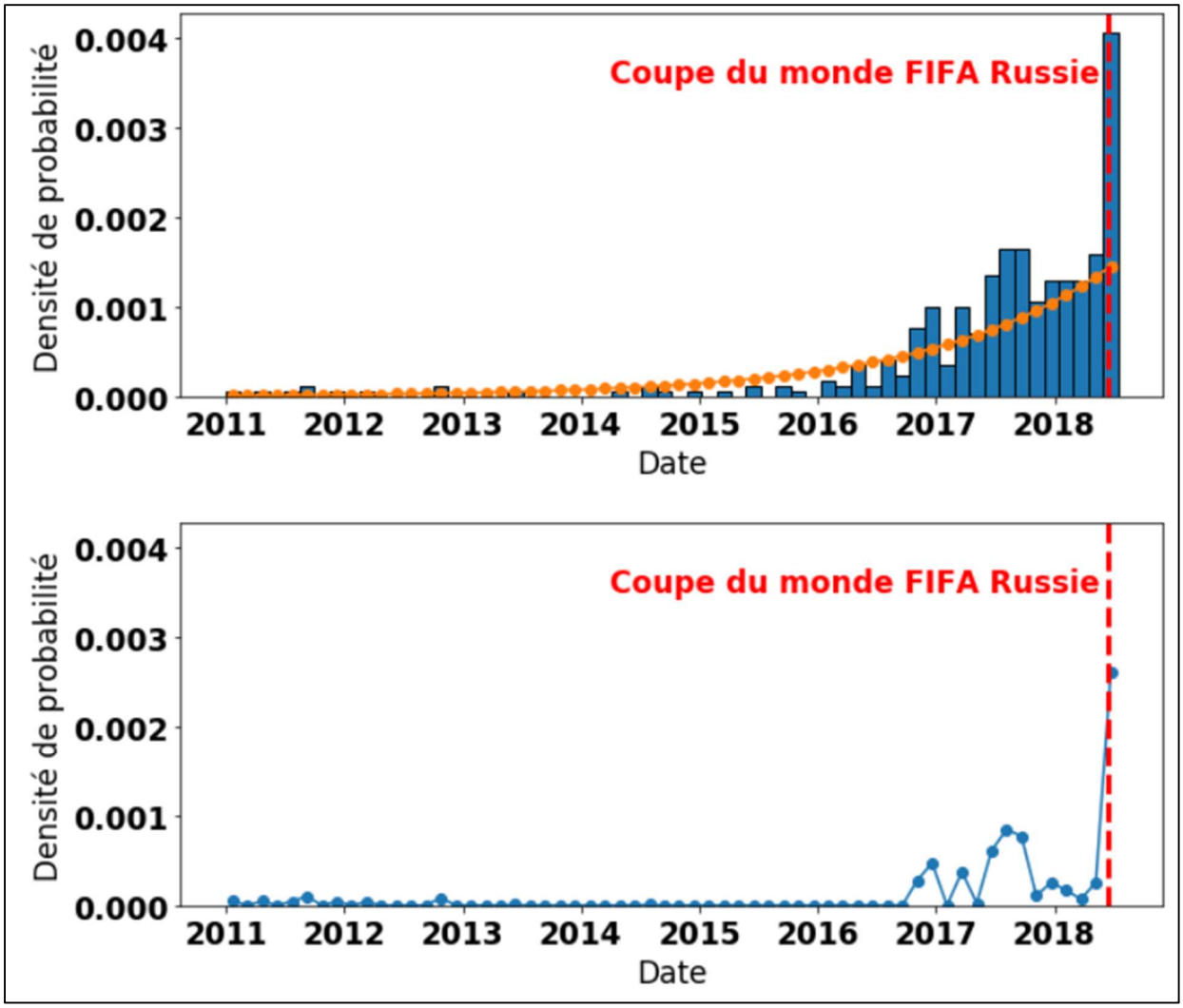

Figure 4 Distribution des tweets de l'amas dont 100% des tweets contiennent le mot world et 20% contiennent le mot **cup**.

Le graphique du haut présente la distribution brute ainsi que la tendance exponentielle (orange) de l'ensemble de la base de données. La figure du bas présente la distribution après le retrait de la tendance de l'ensemble de la base de données. La date identifiée comme **Coupe du monde FIFA Russie** est celle du début de la coupe du monde de football (soccer) 2018 en Russie.

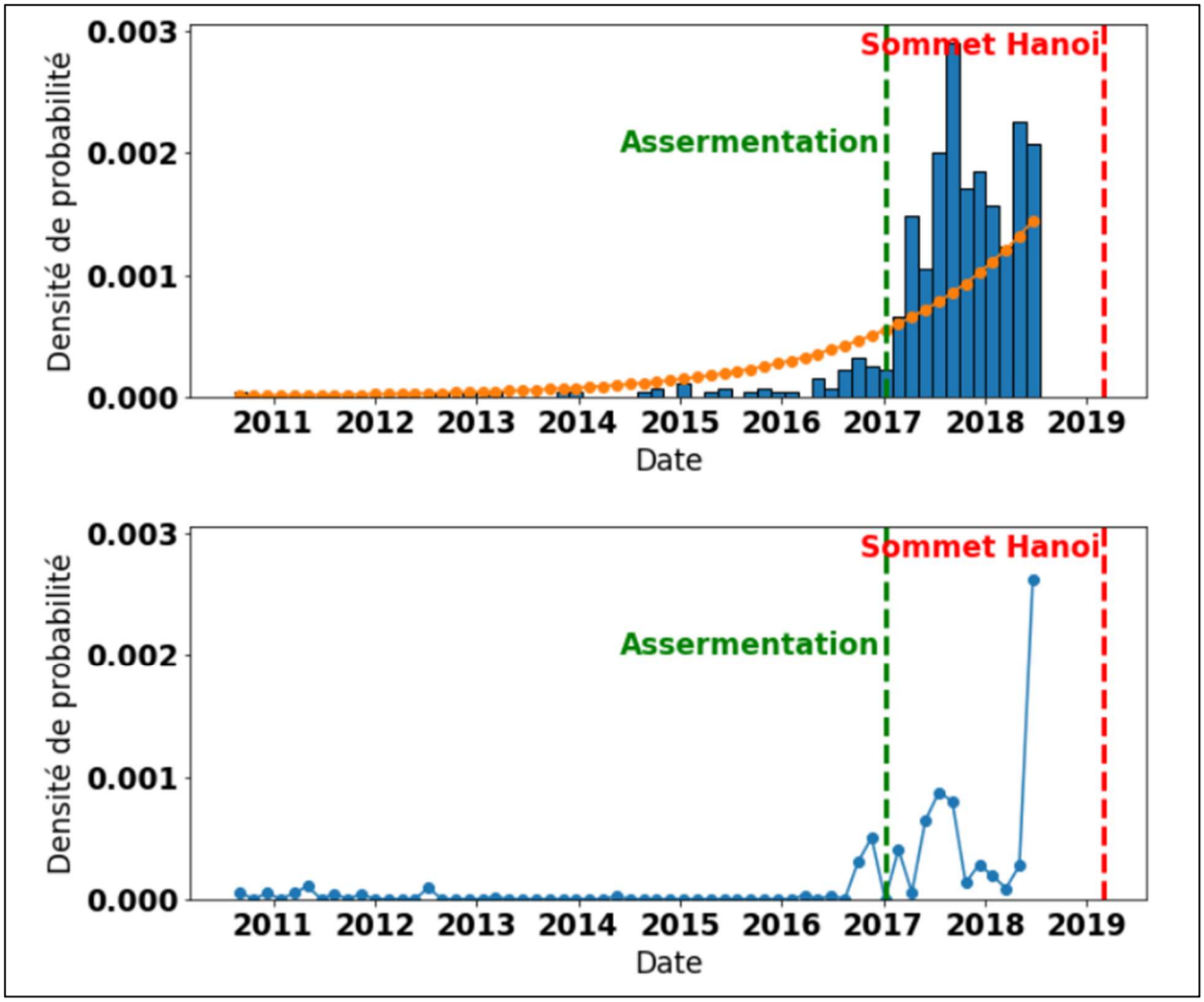

Figure 5 Distribution des tweets de l'amas dont 88% des tweets contiennent le mot **north**, 71% contiennent le mot **korea** et 17% contiennent le mot **south**.

Le graphique du haut présente la distribution brute ainsi que la tendance exponentielle (orange) de l'ensemble de la base de données. La figure du bas présente la distribution après le retrait de la tendance de l'ensemble de la base de données. La date identifiée comme étant **assermentation** représente la date d'entrée en fonction du président Donald Trump. Celle identifiée comme étant **sommet Hanoi** représente la date le rencontre entre Donald Trump et Kim Jong-un. Il serait intéressant d'obtenir des tweets de la fin 2018 et début 2019 pour observer la croissance fulgurante de la densité de probabilité.

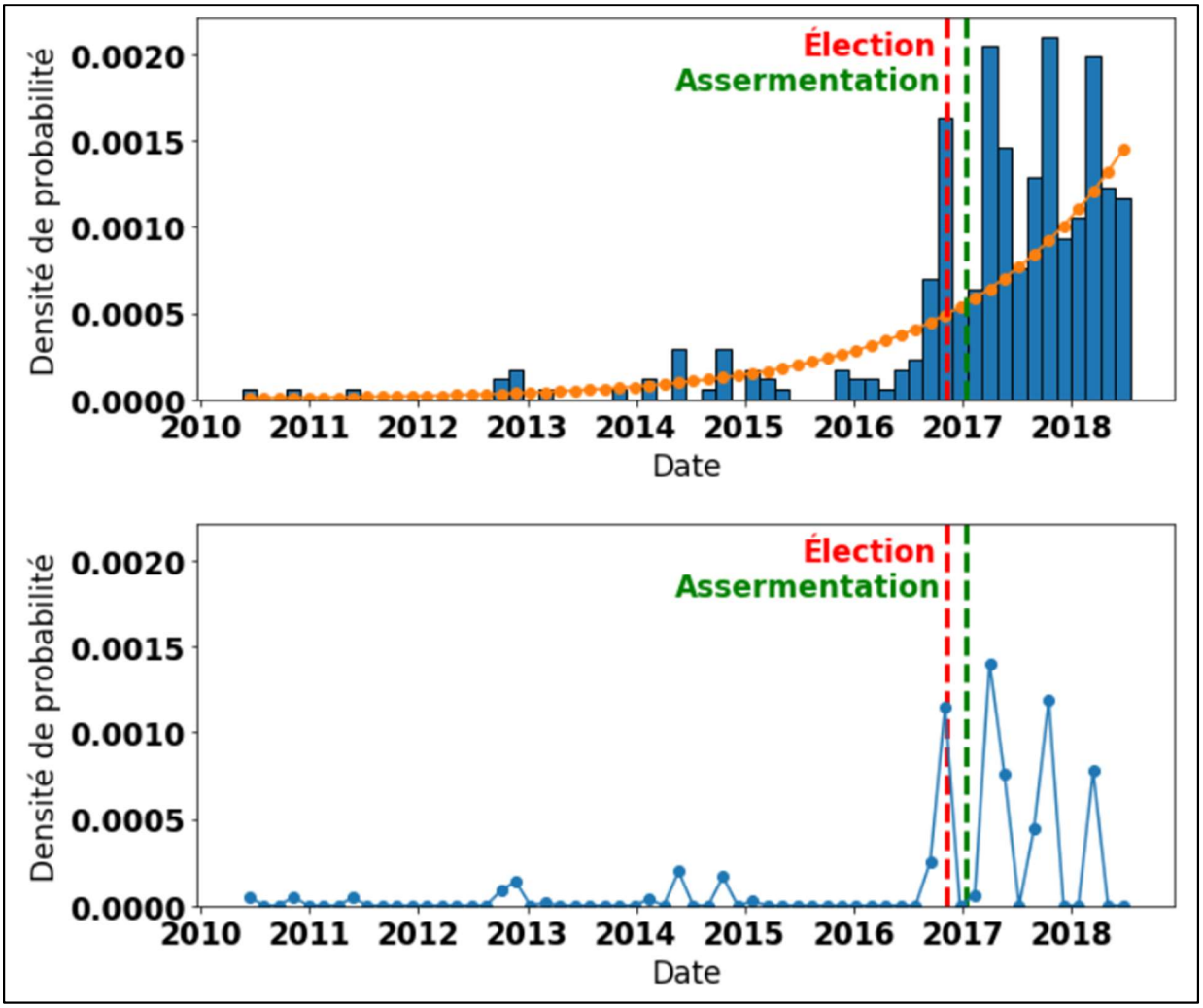

Figure 6 Distribution des tweets de l'amas dont 99% des tweets contiennent le mot **elect** et 18% contiennent le mot **result**.

Le graphique du haut présente la distribution brute ainsi que la tendance exponentielle (orange) de l'ensemble de la base de données. Celui du bas présente la distribution après le retrait de la tendance de l'ensemble de la base de données. La date identifiée comme **Élection** est celle de l'élection de Donald Trump. Celle identifiée comme **Assermentation** est la date d'entrée en fonction de Donald Trump. La densité de probabilité est donc concentrée autour du jour de l'élection et connaît un regain après l'assermentation.

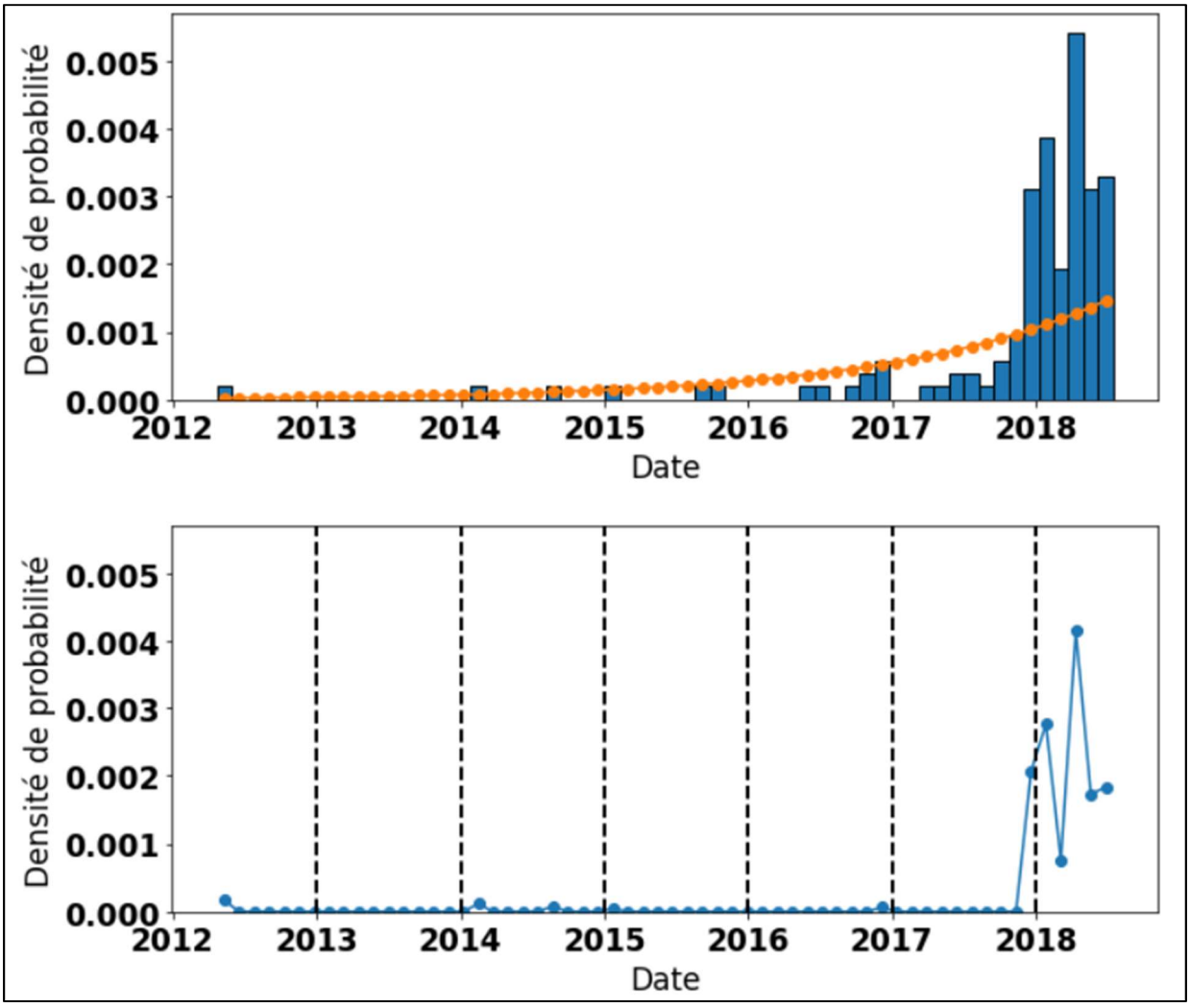

Figure 7 Distribution des tweets de l'amas dont les tweets sont tous du même format, suggérant de la génération automatique de contenu par une machine.

Le graphique du haut présente la distribution brute ainsi que la tendance exponentielle (orange) de l'ensemble de la base de données. Celui du bas présente la distribution après le retrait de la tendance de l'ensemble de la base de données. Cette distribution permet d'estimer la date d'implémentation de ce système de génération automatique de contenu.