



RAPPORT DU PROJET DE FIN D'ÉTUDES

Pour obtenir le diplôme de

Post-Graduation Spécialisée en Big Data et Calcul Intensif

Préparé par :

Adel Guern, Abdesslem Balegh

Analyse des données massives (Big Data) sur les médias sociaux avec Apprentissage Automatique : un intérêt particulier au dialecte Algérien

Soutenu le : 27 MARS 2022

Devant le jury composé de :

Dr. Djamal BELAZOUGUI CERIST Président

Mr. Nadir BOUCHAMA CERIST Examineur

Dr. Abdelbaset KABOU CERIST Encadrant

Analyse des données massives (Big Data) sur les médias sociaux avec Apprentissage Automatique : un intérêt particulier au dialecte Algérien



Adel Guern, Abdesslem Balegh

Mémoire soumis en vue de l'obtention du diplôme de
Post-Graduation Spécialisé

Remerciements

Tous d'abord, nous tenons à remercier le bon Dieu de nous avoir accordé toute la détermination, la volonté et la force pour qu'on puisse réaliser ce modeste travail.

*Nous remercions infiniment notre encadreur **Dr. Abdelbaset KABOU** pour ses conseils, sa patience, sa disponibilité et son soutien tout au long de cette période.*

Nous tenons à remercier également toute l'équipe Big Data du CERIST et notamment Dr. Said YAHIAOUI, Dr. Nadia NOUALI, Mr. Madjid SADALLAH, Mr. Abdelghani KRINAH, Mr. Anis LOUNIS, Mr Adel DEBAH et Mlle Sarra BOUHENNI pour le temps consacré et leurs disponibilité tout au long de cette formation et surtout leurs judicieux conseils, qui ont contribué à alimenter notre réflexion.

Nous tenons à exprimer notre profonde gratitude et nos sincères remerciements aux :

Membres de jury d'avoir accepté de juger notre travail et de l'avoir enrichi.

Nous souhaitons adresser nos remerciements les plus sincères aux personnes qui nous ont apporté leur aide de près ou de loin et qui ont contribué à l'élaboration de ce mémoire ainsi qu'à la réussite de cette formidable période de formation.

Dédicaces

Je dédie ce modeste travail à ma famille avec tous mes sentiments de respect, d'amour, de gratitude et de reconnaissance pour tous les sacrifices déployés pour m'élever dignement et assurer mon éducation dans les meilleurs conditions.

A ma mère et mon père que Dieu ait pitié d'eux et leurs pardonne et leurs habite dans les jardins de la délicité pour l'éducation qu'ils m'ont prodigué ; avec tous les moyens et au prix de toutes les sacrifices qu'ils ont consentis à mon égard, pour le sens du devoir qu'ils mon enseigné depuis mon Enfance.

À ma femme, À mes filles, À mon petit fils. Grace à leur soutient moral tout au long de la période de formation, et leur patience et leur sacrifice de leur temps précieux, de ma profonde tendresse et reconnaissance que Dieu, tout puissant, vous protège et vous garde.

À mes frères, mes soeurs et toute ma famille.

À tous mes amis et mes collègues.

Adel GUERN

Dédicaces

Je dédie ce modeste travail à ma famille avec tous mes sentiments de respect, d'amour, de gratitude et de reconnaissance pour tous les sacrifices déployés pour m'élever dignement et assurer mon éducation dans les meilleurs conditions.

A ma mère que Dieu ait pitié d'elle et lui pardonne et lui habite dans les jardins de la délicité et mon père pour l'éducation qu'ils m'ont prodigué ; avec tous les moyens et au prix de toutes les sacrifices qu'ils ont consentis à mon égard, pour le sens du devoir qu'ils m'ont enseigné depuis mon Enfance.

À ma femme, À ma fille Israa Nour El Houd, À mes fils Adem, Ishak et Ahmed. Grace à leur soutien moral tout au long de la période de formation, et leur patience et leur sacrifice de leur temps précieux, de ma profonde tendresse et reconnaissance que Dieu, tout puissant, vous protège et vous garde.

*À mes frères, ma soeur et toute ma famille.
À tous mes amis et mes collègues.*

Abdesslem BALEGH

ملخص

يعد تحليل وسائل التواصل الاجتماعي ، بما في ذلك تحليل المشاعر المستمدة من بيانات Twitter ، موضوعًا ساخنًا. يهتم بفهم آراء ومشاعر مستخدمي شبكة تويتر فيما يتعلق بالمواضيع المختلفة سواء كانت سياسية ورياضية وفنية وحتى اقتصادية وتجارية. يقع مشروعنا ضمن هذا المنظور. إنه استمرار منطقي لسلسلة من المشاريع التي تهدف إلى إنشاء نظام أساسي قابل للتطوير بدرجة كبيرة لتحليل البيانات الضخمة للمعطيات من وسائل التواصل الاجتماعي. ركز العمل المنجز من قبل بشكل أساسي ، بالإضافة إلى اللغة الإنجليزية ، على معيار اللغة العربية. في هذا الموضوع ، نريد تعزيز هذه المنصة بهدف تحسين إدارة اللهجة الجزائرية ، وهي حاضرة أكثر في سياق اللغة الجزائرية الويب الاجتماعي. خصوصية هذا التحليل ، في حالتنا ، هي أنه يتم إجراؤه عن طريق الدمج كل من أدوات البيانات الضخمة ، بالإضافة إلى تقنيات من عالم التعلم الآلي (تلقائي التعلم). الهدف هو تصميم منصة تسمح تلقائيًا ، وذات صلة ، وقابلة للتطوير بدرجة كبيرة والتصنيف في الوقت الفعلي للتغريدات التي تم تنزيلها أثناء البث. سوف نظهر كيف هذا ممكن من نص بسيط (سقسقة) وكيف نصل إلى مثل هذه النماذج التعليمية باستخدام مجموعة متنوعة من الأدوات والمكتبات مثل Spark ML و Spark NLP وما إلى ذلك.

الكلمات المفتاحية: تويتر ، البيانات الضخمة ، تحليل المشاعر ، التعلم الآلي.

Abstract

Social media analysis, including sentiment analysis drawn from Twitter data, is becoming a highly popular subject. It is interested in understanding the opinions and feelings of users of the Twitter network regarding various topics, whether political, sports, artistic, and even economic and commercial. Our project falls within this perspective. It is a logical continuation of a series of projects aimed at creating a highly scalable platform for big data analysis of data from social media. The work done before has mainly focused, in addition to the English language, on the standard Arabic language. In the current project, we aim to strengthen this platform with an improving the management of the Algerian dialect, which is more present in the context of the Algerian language social web. The peculiarity of this analysis, in our case, is that it is performed by integrating both tools of big data, as well as techniques from the world of machine learning. The goal is to design a platform that allows automatic, relevant, highly scalable and real-time categorization of downloaded tweets. We will show how this is possible from a simple text (tweet) and how to access such models using a variety of tools and libraries such as Spark ML, Spark NLP, etc.

Keywords: Twitter, Big Data, Sentiment Analysis, Machine Learning.

Résumé

L'analyse des médias sociaux, y compris l'analyse des sentiments tirée des données de Twitter, est un sujet populaire. Elle s'intéresse à comprendre les opinions et sentiments des utilisateurs du réseau Twitter sur divers sujets, qu'ils soient politiques, sportifs, artistiques, voire économiques et commerciaux. Notre projet s'inscrit dans cette perspective. Il s'agit d'une suite logique d'une série de projets visant à créer une plate-forme hautement évolutive pour l'analyse des mégadonnées issues des médias sociaux. Le travail effectué auparavant s'est principalement concentré, en plus de la langue anglaise, à la langue arabe standard. Dans le cadre du travail actuel, nous souhaitons renforcer cette plateforme dans le but d'améliorer la gestion du dialecte algérien, plus présent dans le contexte du web social en Algérie. La particularité de cette analyse, dans notre cas, est qu'elle est effectuée en intégrant à la fois des outils de big data, ainsi que des techniques issues du monde de l'apprentissage automatique. L'objectif est de concevoir une plateforme permettant une catégorisation automatique, pertinente, hautement évolutive et en temps réel des tweets téléchargés en Streaming. Nous montrerons comment cela est possible à partir d'un simple texte (tweet) et comment peut-on avoir de tels modèles automatiques en utilisant une variété d'outils et de bibliothèques tels que Spark ML, Spark NLP, etc.

Mots Clefs : Twitter, Big data, Analyse de Sentiments, Apprentissage automatique.

Contents

1	Big Data	13
1.1	Introduction	13
1.2	Définition du Big Data	13
1.3	Domaine de big data	14
1.4	Les caractéristiques du Big Data	14
1.5	Composants d'une architecture Big Data	15
1.6	Les challenges du Big Data	16
1.7	Les acteurs du BIG DATA	17
1.8	Qu'est-ce qu'une plateforme Big Data ?	18
1.8.1	Qu'est-ce qu'Hadoop ?	18
1.8.2	Apache Spark	18
1.8.3	Apache Storm	20
1.8.4	Apache Flink	20
1.8.5	Apache Samza	21
1.9	UN CHAMPS IMMENSE D'APPLICATIONS	21
1.10	Réseaux Sociaux	23
1.10.1	Twitter	23
1.10.2	Contenu des tweets	23
1.10.3	Metadonnées des Tweets et des utilisateurs	23
1.10.4	L'API Twitter	24
1.10.5	Les classes d'API Twitter	25
2	MACHINE LEARNING	26
2.1	Qu'est-ce que l'apprentissage automatique ?	26
2.2	Apprentissage automatique et big data :	26
2.3	Modèle d'apprentissage automatique :	26
2.4	Principaux type d'apprentissage automatique (machine Learning) :	26
2.5	principaux Algorithmes d'apprentissage automatique :	27
2.5.1	Régression linéaire :	27
2.5.2	Régression logistique :	27
2.5.3	Support Vector Machine (SVM) :	27
2.5.4	Classification naïve bayésienne :	27
2.5.5	K-moyennes :	27
2.5.6	K plus proches voisins : [51]	28
2.5.7	réseaux de neurones :	28
2.5.8	L'arbre de décision :	28
2.5.9	Les Forêts Aléatoires : [54]	28
2.6	Evaluation des algorithmes d'apprentissage :	29
2.7	Conclusion :	29
3	Analyse des sentiments	31
3.1	Introduction	31
3.2	La richesse de la langue arabe	31
3.2.1	Eléments de structure de la langue arabe	31
3.2.2	Dialecte Algerien	31
3.3	Définition de l'analyse des sentiments	32
3.4	Catégorisation des sentiments	32
3.5	Algorithmes d'analyse des sentiments	33
3.5.1	Approche automatique	33
3.5.2	Approche à base de règles (Rule-based)	33
3.5.3	Approche hybride	33
3.6	Difficultés de l'analyse de sentiments	34
3.7	Domaines d'Applications	35
3.7.1	La santé	35
3.7.2	Affaires et marketing	36
3.7.3	Ville intelligente	36
3.8	Travaux réalisés en langue arabe	38
3.8.1	Cas du dialecte algérien	38
3.8.2	Cas de dialecte tunisien	38

3.8.3	Cas de dialecte marocain	38
3.9	Conclusion	39
4	Conception, Réalisation et Déploiement	42
4.1	Introduction	42
4.2	architecture	42
4.3	Collecte de tweets	42
4.4	les principales méthodes utilisées en NLP	44
4.4.1	La phase de prétraitement	44
4.4.2	La phase d'apprentissage des données au modèle	46
4.4.3	Extraction des caractéristiques	46
4.4.4	Choix du meilleur modèle ML	47
4.4.5	Fonctionnement de l'analyse des sentiments avec l'apprentissage automatique	47
4.5	Réalisation et Mise en oeuvre	47
4.5.1	Choix d'outils	47
4.6	Déploiement et Configuration	50
4.7	Implémentation	51
4.7.1	Introduction	51
4.7.2	Expérimentations et résultat	51
4.7.3	Résultats de la méthode LEXIQUE :	54
4.7.4	Résultat d'apprentissage supervisé:	55
4.7.5	prétraitement des tweets textes	59
4.7.6	Annalyse sentiments des tweets en streaming	60
4.8	Conclusion	67

List of Figures

1	Comparaison gigabyte et exabyte[48]	13
2	Multiples d’octets[38]	14
3	Les caractéristiques (5V) du Big Data[26]	15
4	Big-data-pipeline[27]	16
5	Quelques acteurs de BIG DATA	17
6	Ecosystème Hadoop[28]	19
7	Cluster Spark : composantes principales.	20
8	Clustering Algeria’s map by dialects	32
9	Workflow de l’analyse des sentiments	33
10	Approches automatiques d’analyse de sentiments reposant sur le ML	34
11	Correspondre avec les données pour déterminer la polarité.	34
12	Etapes du processus proposé pour l’analyse des sentiments.	39
13	Diffèrent résultat des classificateur avec plusieurs configuration.	40
14	Trois étapes majeures de notre architecture	42
15	La chaine de valeur big data	43
16	Processus de collection des tweets utilisant Twitter API[40]	44
17	Authentification et accès au tweets utilisant Twitter4j[46]	44
18	Representation des vecteurs issues de la methode Term-Frequency TF[41]	45
19	Pipeline d’analyse de sentiments en utilisant un modèle ML[42]	47
20	Outils et Composantes de base de notre Architecture.	48
21	Spark streaming.	48
22	Spark ML	49
23	Exemple du pipeline d’analyse de sentiments en utilisant un modèle ML (logistic regression dans ce cas).	49
24	Cluster Hadoop de notre solution.	50

List of Tables

1	Table 1.1 – Comparaison des différents outils Big Data [47]	22
2	Distribution des données collectées selon leurs thèmes.	38
3	Statistiques de corpus TSAC.	38
4	Résultats d’expériences d’Analyse de Sentiment tunisien en utilisant divers classificateurs	38
5	Exemple de prétraitement d’un commentaire.	39
6	Nombre de commentaires par polarité.	50
7	Exemple de notation de quelques commentaires.	50
8	Caractéristiques techniques des machines utilisées dans le déploiement.	51
9	Principaux fichiers à modifier pour la configuration de notre infrastructure.	51
10	Évaluation de performance des différents algorithmes utilisés sur le corpus testé (NumFeatures(100))	57
11	Évaluation de performance des différents algorithmes utilisés sur le corpus testé (NumFeatures(200))	57
12	Évaluation de performance des différents algorithmes utilisés sur le corpus testé (NumFeatures(1000))	57
13	Évaluation de performance des différents algorithmes utilisés sur le corpus testé (NumFeatures(2000))	58
14	Évaluation de performance des différents algorithmes utilisés sur le corpus testé (NumFeatures(3000))	58
15	Évaluation de performance des différents algorithmes utilisés sur le corpus testé (NumFeatures(5000)).	58
16	Évaluation de performance des différents algorithmes utilisés sur le corpus testé (NumFeatures(10000))	58
17	Évaluation de performance des différents algorithmes utilisés sur le corpus testé (NumFeatures(15000))	58
18	Évaluation de performance des différents algorithmes utilisés sur le corpus testé (NumFeatures(20000))	59