REPUBLIQUE ALGERIENNE DEMOCRATIQUE & POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR & DE LA RECHERCHE SCIENTIFIQUE

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

Soutenue le : 27 MARS 2022

Devant le jury composé de :

Dr. Djamal BELAZOUGUI CERIST Président

Mr. Nadir BOUCHAMA CERIST Examinateur

Dr. Abdelbaset KABOU CERIST Encadrant

Année Universitaire : **2019 - 2020**

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

Alger Fevrier 2022

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 mon enseigné depuis mon Enfance.

À ma femme, À ma fille Israa Nour El Houd, À mes fils Adem, Ishak et Ahmed. 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, 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

List of Figures

List of Tables