République Algérienne Démocratique et Populaire
الجمهورية الجزائرية الديمقراطية الشعبية
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
وزارة التعليم العالي و البحث العلمي

ECOLE NATIONALE SUPÉRIEURE D'INFORMATIQUE

المدرسة الوطنية العليا للإعلام الآلي
(المعهد الوطني للتكوين في الإعلام الآلي سابقا)
Ecole nationale Supérieure d'Informatique
ex. INI (Institut National de formation en Informatique)

Mémoire de fin d'étude

Pour l'obtention du diplôme d'ingénieur d'Etat en Informatique

Option : Systèmes d'Information et Technologie

Thème

Classification et Analyse de tonalité des sujets de l'actualité

Encadré par

- Pr CHALAL Rachid (ESI)
- Dr BOUSBIA Nabila (ESI)
- Dr AMRANE Abdesslam (CERIST)

Réalisé par

- KHORSI Roufaida
- YEBDA Sadia

Promotion 2018/2019

Dédicaces

A ma famille

Elle qui m'a doté d'une éducation digne, son amour fait de moi ce que je suis aujourd'hui

A mon binôme Roufaida

Pour les moments agréables et difficiles que nous avons passé durant cette année

A tous mes amis, tout particulièrement Amel

A toute personne qui m'a supporté de près ou de loin

Sadia

Dédicaces

A mes très chers parents,

Qui n'ont jamais cessé de croire en moi, ils m'ont doté d'une éducation digne et leur amour fait de moi ce que je suis aujourd'hui.

A ma sœur Asma et mon frère Tahar

Qui ont toujours su trouver les mots pour me soutenir.

A mon binôme Sadia

Pour cette agréable expérience que nous avons vécue cette année.

À tous ceux qui sont proches et dont je n'ai pas cité le nom.

Roufaida

Remerciements

Nous tenons à remercier toutes les personnes qui ont contribué au bon déroulement de notre projet de fin d'études.

De prime abord, nous adressons nos vifs remerciements à nos encadrants, Mme BOUSBIA Nabila, et Mr CHALAL Rachid ; pour leur disponibilité, leur aide précieuse ainsi que leurs conseils judicieux.

Nous remercions également notre promoteur Mr AMRANE Abdesslam au niveau du CERIST pour son accompagnement et ses explications qui nous ont été précieux dans l'aide apportée.

Nous tenons à remercier toute la famille ESI, passant du directeur de l'école jusqu'au simple travailleur pour leur engagement et effort durant notre période d'études. En particulier, Mme. AIT ALI YAHIA Dahbia, pour sa disponibilité afin de pouvoir réaliser notre PFE dans les meilleures conditions.

Nous présentons nos respects et remerciements aux membres du jury qui ont accepté d'évaluer notre travail.

Nos remerciements s'étendent également à tous les enseignants qui nous ont transmis leur savoir, partagé leur expérience et qui par leurs compétences nous ont soutenu dans la poursuite de nos études.

Résumé

Les sources de média sont considérées comme la première référence d'information. Les technologies de diffusion de masse produisent en permanence une quantité énorme d'informations et attirent un large public. Souvent, ces médias jouent à la fois un rôle de miroir d'opinion de la société et de créateur d'opinion. La mise en évidence de certaines informations peut avoir un impact économique et politique pour les organisations.

La catégorisation manuelle de ces informations demeure une tâche difficile aux agents de veille. Ceci a suscité un intérêt particulier pour l'automatisation de la catégorisation des textes véhiculés dans l'actualité.

Il existe plusieurs travaux dans le cadre de la classification de texte par apprentissage automatique qui ont été réalisés dans diverses langues. Ce projet vise les sources média algériennes et traite le texte exprimé en langue française. L'objectif principal de ce travail est de développer un système de classification de l'actualité basé sur l'extraction de texte pour classifier automatiquement les sujets de l'actualité et analyser la tonalité exprimée.

Ce document présente en premier lieu un état de l'art sur les différentes approches de classification de texte ainsi que les approches d'analyse de tonalité basées sur l'apprentissage automatique. L'étiquette utilisée pour les catégories est l'ensemble des entités nommées désignant des organisations et/ou leur secteur d'activité, tandis que les étiquettes d'opinion indiqueront l'orientation positive, négative ou neutre des informations vers l'entité nommée. Deux modèles de classification ont été conçus puis évalués sur une série de tests effectués sur un dataset que nous avons collecté. Les modèles proposés atteignent des performances très élevées en termes de qualité de prédiction et d'erreurs d'estimation.

Enfin, et pour de montrer l'intérêt pratique de notre solution nous présentons un prototype qui illustre un cas d'utilisation du système. L'utilisateur introduit ses articles et l'application retourne leur classe thématique et leur tonalité.

Mots clés *: classification de texte, Fouille d'opinion, Traitement automatique de langage naturel, Apprentissage automatique, Veille média.*

Abstract

Media sources are considered as the first information reference. Mass broadcasting technologies produce continuously huge amount of information and acquire a widespread audience. As a result, it can easily guide people's attention by highlighting some specific news. One of the main challenges of organizations currently, is to keep an eye on all media sources in order to track the success of their releases, manage their reputation and know their industry and competitors.

Since organizations aim to stay up-to-date with the latest trends and to transform media news into knowledge for decision makers, daily media debrief organized by topic is needed. But manual categorization of a large broadcasted news may sound time and effort consuming task. That is why there is a particular interest for making texts categorization automatic.

In practice, several methods in the context of automatic text classification have been used. The most famous is machine learning which has proved its performance in many languages. Our project aims to monitor Algerian media sources and analyze the French texts produced. The main objective of this work is to develop an accurate news classification system which applies text mining to automatically classify news by topics and by tonality.

Within this context, we provide in this paper a state of the art related to text classification and tone analysis. We also propose a news text classification tool that categorizes automatically media news by topic and by tone. In our system two classification models are designed and evaluated with series of tests performed on a large dataset (over 74.000 article). The proposed models achieved very high performance in terms of quality of prediction, error estimation and processing time.

***Key Words**: Text categorization, Opinion mining, Natural language Processing, Machine learning, Media monitoring.*

ملخّص

تعتبر وسائل الإعلام المرجع الأول للمعلومات. حيث تنتج تقنيات البث الشامل باستمرار كمية هائلة من المعلومات وتجذب جمهورًا واسعًا. غالبًا ما تلعب هذه الوسائل دورًا معاكسًا للمجتمع ومبدع للرأي. يمكن أن يكون للفت الانتباه وتسليط الضوء على بعض المعلومات تأثير اقتصادي وسياسي على المنضمات.

يظل التصنيف اليدوي لهذه المعلومات مهمة صعبة عند مراقبة وسائل الإعلام. وقد أثار هذا اهتمامًا خاصًا لجعل تصنيف النصوص المنقولة في الأخبار اوتوماتيكيا.

هناك العديد من الأعمال والأبحاث تم القيام بها في هذا السياق وبلغات مختلفة. يهتم هذا المشروع بمصادر إعلامية جزائرية ويتعامل مع النص المكتوب باللغة الفرنسية. الهدف الرئيسي لهذا البحث هو تطوير نظام تصنيف للأخبار بالاعتماد على استخراج النص لتصنيف مواضيع الأخبار تلقائيًا وتحليل الرأي المعبر عنه في النص.

تقدم هذه مذكرة أولاً الأساليب والتقنيات المطوّرة لتصنيف النص إلى فئات وكذلك مناهج تحليل الرأي القائمة على التعلم الآلي. التسمية المستخدمة لتحديد الفئة الموضوعيّة للنص هي أسماء المنظمات او الشركات، القطاعات الاقتصادية او محيط نشاط الشركة، في حين أن ملصقات الرأي هي ثلاثة: إيجابي، سلبي أو حيادي.

تم تصميم نموذجين للتصنيف وتقييمهما على سلسلة من الاختبارات التي أجريت على مجموعة كبيرة من النصوص (74.000 نص) . حققت النماذج المقترحة أداءً عاليًا من حيث جودة دقة التنبؤ والتقدير.

الكلمات المفتاحية: *تصنيف النص، تحليل الرأي، معالجة اللغة، التعلم الآلي، مراقبة وسائل الاعلام.*

Table des matières