عنوان الرسالة: تحليل وتصنيف محتوى المدونات المصغرة: الاستفادة من المحتوى الناتج عن مستخدميها لخدمة مجال الأعمال

اسم الطالب: نوره علي العثمان

اسم المشرف: د. شيماء سلامة

المستخلص العربي: نظرا لأن المدونات المصغرة أصبحت من المصادر الأساسية للمعلومات على الويب، فقد جذبت دراسة هذه المعلومات الاهتمام الشديد للعديد من الأطراف مثل العملاء والشركات والمؤسسات. وذلك بسبب أن محتوى المدونات المصغرة يمكّنهم من استكشاف آراء عامة الناس حول أي موضوع معين. تشمل الموضوعات ذات الأهمية تصنيف محتوى المدونات المصغرة وكشف الأخبار وتحليل الإهتمامات والتنقيب عن الآراء. في مجال تصنيف محتوى المدونات المصغرة، الدراسات الحالية شاملة نسبيا. ومع ذلك، يظهر المسح الأدبي نقصا واضحا في الدراسات التي تأخذ في عين الأعتبار الدور المهم للمحتوى الذي ينشئه المستخدمون وخاصة باللغة العربية، نظرا لطبيعتها اللغوية الفريدة من حيث التهجئة والصرف واللهجات مما يجعل المهمة أكثر صعوبة. لهذه الأسباب، تقدم هذه الرسالة مقترح لتحسين تصنيف محتوى اللغة العربية في المدونات المصغرة. الخطوة الأولى تتمثل في تحسين الأدوات التي تساعد في عملية التحليل من خلال بناء مصادر جديدة للمحتوى العربي. بينما نقترح في الخطوة الثانية، والتي تمثل المساهمة الرئيسية، نموذجا جديدا للتعلم العميق لحل مشاكل تصنيف المحتوى واهتمامات المستخدمين المختلفة.أظهرت التجارب المكثفة أن نموذجنا تعامل مع المحتوى ومشاكل تصنيفه وتحليله لجميع مدخلات البيانات المختبرة بشكل أفضل بكثير من عدة نماذج أساسية تم اختبارها.

Thesis Title: Content Analysis and Classification on Microblogging Domain: Leveraging User-generated Content for Business.

Student Name: Nourah Ali Alothman

Supervisor Name: Dr. Shaima Salama

English Abstract: As Microblogs become the fundamental source of information on the web, investigating this information diffusion has caught the intense interest of several parties such as customers, companies, business world, and many others. Microblogs content enables them to explore people's opinion about any given topic. Topics of interest involve microblogs classification, news detection, interest analysis, and opinion mining. In microblogs content classification, the current literature is relatively extensive. However, the literature shows an apparent lack of studies that consider the significant role of user-generated content and preferences. In particular, for the Arabic language, posing unique linguistic complexities in terms of orthography, morphology and dialects, makes the task more challenging. Therefore, this work provides the rationale behind the existing work and propose methods to enhance microblogs content classification in the Arabic language. The first step is to improve the tools that aid in the process of analysis; providing an annotated Arabic dataset is the most critical aspect of this process. For the English language, many free corpora are available, and yet these tools are also scarce in other languages, such as Arabic. This research describes the author's work to enrich user-interest classification in Arabic by building a new Arabic interest-based Corpus. The second step, representing the main contribution, proposes a new hybrid deep learning model – Deep Neural Network with Gradient Boosting (DeepGB) for classification problems. To learn the input features, DeepGB consists of multiple stacked layers that will eventually learn features, followed by the Gradient Boosting classifier in the last layer to predict class labels. The conducted experiments showed that the proposed model reached almost 97\% f1-score, and handled the classification problems for all the tested data significantly better than several baseline models.