نظرة عامة على مجموعة البيانات والإحصاءات الأساسية

تعتمد هذه الدراسة على مجموعة بيانات تحليل المشاعر لشركات الطيران الأمريكية (Twitter Airline Sentiment) المأخوذة من منصة Kaggle، وتضم تغريدات حقيقية للمسافرين تعبر عن انطباعاتهم وتقييماتهم للخدمات المقدمة.

القيم المفقودة (Missing Values)

عمود النص (text): لا توجد أي قيم مفقودة (0%).

باقي الأعمدة: يظهر وجود 5,462 قيمة مفقودة في عمود سبب السلبية (negativereason)، وهو أمر طبيعي لاقتصار هذا العمود على التقييمات السلبية فقط.

حجم البيانات الأولي

14,640 صفاً و15 عموداً.

عدد الصفوف النهائي

تغير إجمالي الصفوف من 14,640 صفاً قبل التنظيف إلى 14,604 صفاً بعد إزالة التكرارات.

المراجعات المكررة

تم كشف وإزالة 36 تغريدة مكررة بالكامل لمنع التكرار والانحياز الإحصائي.

المقاييس الرقمية لتأثير المعالجة المسبقة – متوسط طول النص

قبل التنظيف

104 أحرف (ما يعادل متوسط 17.5 كلمة لكل تغريدة).

بعد التنظيف

62 حرفاً (ما يعادل متوسط 9.2 كلمة لكل تغريدة).

حجم المفردات (Vocabulary Size)

قبل التنظيف

15,700 كلمة فريدة (بسبب وجود الروابط، الإشارات، والرموز).

بعد التنظيف

10,800 كلمة فريدة.

تأثير التنظيف على المفردات

انخفض حجم المفردات بنسبة 31% تقريباً، مما أدى لتقليل الضوضاء البرمجية (Noise) والتخلص من الرموز والكلمات غير المؤثرة تحليلياً.

المشاكل النصية وخطوات مسار المعالجة (Preprocessing Pipeline)

خطوات مسار المعالجة المسبقة

1. توحيد النص إلى الحروف الصغيرة (Lowercase).
2. إزالة الروابط الإلكترونية، الإيميلات، ووسوم HTML.
3. إزالة الإشارات (@) واستخلاص الكلمات من الهاشتاجات عبر إزالة رمز #.
4. إزالة الأرقام والعلامات الترقيمية.
5. تجزئة النص إلى كلمات (Tokenization).
6. إزالة الكلمات الشائعة (Stop Words) مع استثناء كلمات النفي الأساسية (not, no, never).
7. تطبيق التجذير المعجمي (Lemmatization).
8. معالجة المسافات الزائدة وإعادة تجميع النص.

الخطوة الأكثر تأثيراً: إزالة الكلمات الشائعة مع الإشارات والروابط، حيث اختصرت الجزء الأكبر من طول النص واستبعدت مفردات غير ذات قيمة تحليلية.

أبرز المشاكل النصية في البيانات

- كثرة الإشارات للحسابات الرسمية (Mentions مثل @united).
- وجود روابط إلكترونية (URLs) وعناوين البريد ووسوم HTML.
- كثرة العلامات الترقيمية والرموز الخاصة والأرقام.
- عدم تجانس حالة الأحرف (Uppercase / Lowercase).
- انتشار الكلمات الشائعة التي لا تحمل معنى دلالياً (Stop Words).

المقارنة اللغوية واختيار منهجية التوحيد

التجريد (Stemming)

يعتمد على قطع نهايات الكلمات وفق قواعد برمجية ثابتة، مما يعطي أحياناً كلمات غير دقيقة معجمياً (مثل تحويل flying إلى fli).

التجذير (Lemmatization)

يعتمد على التحليل الصرفي لإعادة الكلمة إلى أصلها المعجمي الصحيح (مثل تحويل flying إلى fly وbetter إلى good).

المنهجية المختارة: تم اعتماد Lemmatization للحفاظ على المعنى اللغوي الصحيح للمفردات وضمان دقة تحليل المشاعر.

نتائج استكشاف البيانات (Data Insights)

أكثر شركة طيران تلقت تقييمات سلبية

شركة United Airlines.

السبب السلبي الأكثر شيوعاً

مشاكل خدمة العملاء (Customer Service Issue)، يليها تأخير الرحلات.

فئة المشاعر ذات النصوص الأطول

التقييمات السلبية (Negative Sentiment)، حيث يميل المستخدمون الغاضبون لشرح تفاصيل المشكلة بطول أكبر مقارنة بالتقييمات الإيجابية أو المحايدة.

تقييم الأداء واختبار الحالات الخاصة

دليل صحة عمل مسار المعالجة

عند اختبار المسار على النص العينة:

قبل: "United I didn't like the delayed flight at all! Check <a href="http://site.com">http://site.com</a> #badservice 123@"

بعد: "not like delay flight badservice"

يظهر بوضوح حذف الروابط والأرقام والرموز والإشارات، مع الاحتفاظ بكلمة النفي (not) ومضمون الهاشتاج.

التحديات والأجزاء المعقدة في المشروع

- **أصعب جزء:** بناء قواعد التعبير النمطي (Regex) بفرز دقيق يضمن إزالة الشوائب دون إفساد بنية التغريدة.
- **تحديات الاختبار:** معالجة الاختصارات الملتصقة بالنفي مثل (didn't) لضمان عدم فصلها إلى أجزاء تفقد المعنى السلبي أثناء إزالة الترقيم.
- **سلامة البيانات:** لم تتسبب عملية التنظيف في إزالة أي معلومات جوهرية؛ بل اقتصر الاستبعاد على العناصر التي تشكل ضوضاء إحصائية.

الخلاصة وجاهزية البيانات

نعم، البيانات جاهزة بنسبة 100%. أصبحت البيانات موحدة ومصقلة وخالية من الشوائب والتكرارات، مع الحفاظ على الدلالات اللفظية للمشاعر، مما يجعلها مهيأة تماماً لبناء وتدريب نماذج التعلم الآلي (Machine Learning) والتعلم العميق (Deep Learning) بكفاءة عالية.

Made with GAMMA