**Calcul du Kappa de Fleiss pour l’Annotation des Émotions dans un Corpus de Tweets**

**Contexte et Objectif**

Dans le cadre de notre TER réalisé en groupe, nous avons effectué une annotation manuelle des émotions exprimées dans un corpus de 1140 tweets, annotés par trois évaluateurs ***(Alain, Anton, Tony)***. Chaque évaluateur a attribué jusqu’à deux émotions par tweet ***(émotion 1 et émotion 2)*** parmi sept catégories possibles :

* 1= peur
* 2 = colère
* 3 = joie
* 4 = surprise
* 5 = tristesse
* 6 = dégoût
* 7 = neutre

Notre objectif était de mesurer la fiabilité de ces annotations en calculant le **Kappa de Fleiss**, un indicateur de l’accord inter-évaluateurs, avant de définir des annotations définitives pour chaque tweet.

**Méthodologie**

* **Préparation des données** :
  + Nous avons enregistré les annotations dans un fichier CSV avec 15 colonnes, incluant le texte des tweets et les annotations des évaluateurs (par exemple, Alain\_émotions, Alain\_émotions 2, etc.).
  + Un problème de décalage dans les colonnes a été identifié, causé par des virgules et guillemets dans le texte des tweets, perturbant le parsing du fichier. Ce problème a été résolu en remplaçant les virgules par des points (seulement pour le calcul) et un script de nettoyage pour s’assurer que chaque ligne avait exactement 15 colonnes.
* **Calcul du Kappa de Fleiss** :
  + Nous avons calculé le Kappa de Fleiss sous trois approches principales :
    - **Version 1 : Toutes émotions (6 votes par tweet)** : Chaque évaluateur pouvait donner jusqu’à 2 votes par tweet, soit un maximum de 6 votes par tweet (3 évaluateurs × 2 émotions). Cependant, les colonnes pour l’émotion 2 étaient souvent vides, ce qui rendait le nombre de votes inconstant.
    - **Version 2 : Émotion 1 uniquement** : Chaque évaluateur donnait un seul vote (émotion 1), soit exactement 3 votes par tweet.
    - **Version 1 bis : Un vote par évaluateur** : Pour contourner le problème des votes manquants, nous avons combiné émotion 1 et émotion 2 en un seul vote par évaluateur, en donnant la priorité à l’émotion 1.
  + Une approche optimisée a ensuite été développée pour maximiser l’accord inter-évaluateurs :
    - **Version optimisée :** 
      * **Maximiser l’accord avec toutes les émotions** : Pour chaque évaluateur, nous avons choisi entre son émotion 1 et son émotion 2 en fonction de celle qui est en meilleur accord avec **toutes les émotions** des autres évaluateurs (c’est-à-dire émotion 1 et émotion 2 d’Anton et Tony pour Alain, par exemple). Cela prend en compte jusqu’à 6 votes par tweet pour maximiser le consensus.
  + Le calcul a été effectué avec Python, en utilisant les bibliothèques pandas et statsmodels.

**Résultats**

* **Version 1 : Toutes émotions (6 votes par tweet)** :
  + Aucun tweet n’avait exactement 6 votes, car les colonnes pour l’émotion 2 étaient souvent vides (par exemple, Alain\_émotions 2 manquante dans le tweet 0). Résultat : 1140 tweets exclus, aucun Kappa calculé.
  + Cette approche n’était pas viable étant donné la structure des données.
* **Version 1 bis : Un vote par évaluateur (priorité à émotion 1)** :
  + En prenant un vote par évaluateur (priorité à l’émotion 1), le Kappa de Fleiss obtenu est **0.3083**.
  + Selon les seuils de Landis & Koch (1977), cela indique un **accord léger** (0.21–0.40).
* **Version 2 : Émotion 1 uniquement** :
  + En utilisant uniquement l’émotion 1 (3 votes par tweet), le Kappa de Fleiss est également **0.3083**, ce qui est cohérent avec la version 1 bis, car l’émotion 2 n’était pas utilisée dans cette approche.
  + Cela confirme un **accord léger** entre les évaluateurs.
* **Version optimisée : Maximiser l’accord avec toutes les émotions** :
  + En comparant chaque émotion (1 et 2) d’un évaluateur avec toutes les émotions (1 et 2) des autres évaluateurs, le Kappa de Fleiss obtenu est **0.3758**.
  + Cela représente une amélioration par rapport au Kappa initial de 0.3083, montrant que l’émotion 2, lorsqu’elle est présente, peut refléter un consensus dans certains cas où l’émotion 1 diverge. Selon les seuils de Landis & Koch, ce Kappa indique toujours un **accord léger** (0.21–0.40), mais il se rapproche de la limite de l’accord modéré (0.41–0.60).
* **Annotations définitives** :
  + Nous avons généré un fichier resultats\_avec\_emotion\_majoritaire.csv, contenant trois colonnes :
    - emotion\_majoritaire\_toutes : Émotion majoritaire basée sur toutes les émotions (peu fiable ici à cause des données manquantes).
    - emotion\_majoritaire\_em1 : Émotion majoritaire basée sur l’émotion 1 uniquement.
    - emotion\_majoritaire\_optimisée : Émotion majoritaire basée sur l’approche optimisée, qui maximise l’accord en prenant en compte toutes les émotions.
  + Pour les annotations définitives, nous avons choisi une approche hybride plutôt que de nous baser uniquement sur la colonne emotion\_majoritaire\_optimisée. Nous avons d’abord identifié les tweets où au moins deux évaluateurs étaient en accord (par exemple, votes [3, 3, 5]). Pour ces tweets, nous avons conservé l’émotion majoritaire. Ensuite, pour les tweets restants, nous avons examiné à la fois les émotions 1 et 2 pour trouver un consensus, en combinant les deux si nécessaire. Enfin, pour les cas où aucun accord n’était possible (par exemple, votes [3, 5, 7]), nous nous sommes réunis pour relire chaque tweet et décider collectivement de l’émotion finale.

**Discussion**

* **Niveau d’accord** :
  + Le Kappa initial de 0.3083 (émotion 1 uniquement) indique un accord léger, suggérant que les évaluateurs divergent souvent dans leurs annotations. Cela peut être dû à l’ambiguïté des tweets, à des définitions d’émotions mal comprises, ou à des biais individuels.
  + L’approche optimisée, en prenant en compte toutes les émotions (jusqu’à 6 votes par tweet), a permis d’augmenter le Kappa à **0.3758**, montrant que l’émotion 2 contient des informations utiles pour améliorer l’accord dans certains cas. Bien que l’accord reste léger, il se rapproche de l’accord modéré, ce qui est encourageant pour une première phase d’annotation.
* **Problèmes rencontrés** :
  + Les décalages dans le fichier CSV ont nécessité un nettoyage préalable.
  + L’inconstance du nombre de votes pour la version "toutes émotions" a empêché le calcul initial du Kappa. L’approche optimisée a permis de contourner ce problème en sélectionnant un vote par évaluateur tout en maximisant l’accord avec toutes les émotions disponibles.
* **Limites** :
  + L’émotion 2 est souvent absente (par exemple, dans les 28 premiers tweets, seulement 11 ont au moins une émotion 2), ce qui limite l’impact de cette approche.
  + Les cas où les évaluateurs divergent fortement (par exemple, votes [3, 5, 7]) ont nécessité une résolution manuelle.

**Recommandations et Prochaines Étapes**

* **Amélioration de l’accord** :
  + Nous avons organisé une discussion entre les évaluateurs pour clarifier les définitions des émotions et revoir les tweets où ils divergeaient, notamment ceux sans majorité claire (par exemple, votes [3, 5, 7]). Cette discussion a permis de mieux aligner nos interprétations et de réduire les ambiguïtés.
  + Nous avons également ré-annoté un sous-ensemble des tweets après avoir ajusté les instructions, ce qui a contribué à améliorer la cohérence des annotations.
* **Annotations définitives** :
  + Nous avons résolu manuellement les cas ambigus (votes sans majorité) lors d’une réunion avec les évaluateurs. Pour ces tweets, nous avons relu chaque message ensemble et décidé collectivement de l’émotion finale, en prenant en compte à la fois les émotions 1 et 2 pour refléter au mieux le contenu du tweet.

**Conclusion**

Le calcul du Kappa de Fleiss a révélé un accord initial léger (0.3083) entre les évaluateurs, qui a été amélioré à **0.3758** grâce à une approche optimisée prenant en compte toutes les émotions (1 et 2) de chaque évaluateur. Cela montre que l’émotion 2, bien que moins fréquemment utilisée, peut contribuer à un meilleur consensus dans certains cas. Pour les annotations définitives, nous avons adopté une approche hybride : nous avons conservé les émotions majoritaires lorsque deux évaluateurs étaient en accord, combiné les émotions 1 et 2 pour trouver un consensus sur les tweets restants, et résolu les cas ambigus lors d’une réunion collective. Ce processus nous a permis d’obtenir des annotations définitives fiables, qui serviront de base pour approfondir l’analyse des émotions dans les tweets.

**Références**

* Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

