L'évaluation et les tests

7.1 Les tests des classifications de notation des sentiments

7.1.1 Le choix de validation des performances

Dans le chapitre précédent nous avons présenté les techniques de classification de texte pour effectuer la notation de l'opinion des critiques cinématographiques. Dans ce chapitre nous présentons nos résultats et nous précisons les avantages et les inconvenants des méthodes décrites dans le chapitre précédent. Pour mesurer les performances des classificateurs nous calculons les paramètres du rappel et de la précision, en déduisant la valeur de F-score.

Dans notre activité de recherche nous avons utilisé le classificateur linguistique pour lequel nous avons créé les grammaires en se basant sur les critiques de la base d'apprentissage (identique pour toutes les méthodes). Pour cette raison nous avons choisi la méthode de **validation par test**. Les autres méthodes de validation [*Section 2.3*] sont basées sur l'estimation de l'erreur et utilisent les données de la base d'apprentissage. La création des grammaires pour le classificateur linguistique est basée sur la base d'apprentissage, donc pour calculer la performance nous avons besoin d'une nouvelle base, la base de test. Nous precisons aussi que nous avons un très important nombre de critiques annotées dans notre base de données ce qui justifie l'utilisation de la méthode de validation par test. Nous comparons les résultats de toutes les approches de classification developpées sur le même ensemble de validation.

Nous avons utilisé la même base de test et la même base d'apprentissage pour tous les classificateurs des sentiments. Nous supposons que l'utilisation des mêmes bases d'apprentissage et de tests nous permet d'effectuer la comparaison des résultats des trois classificateurs, même si l'apprentissage était effectué d'une manière complètement différente.

Dans notre recherche de la notation des sentiments, une des méthode utilisée est la méthode de classification de comportement des groupes. Ce classificateur attribue uniquement la note directement à la critique entière. Les autres classificateurs attribuent la note à chaque phrase de la critique. Pour pouvoir comparer les trois méthodes utilisées, la performance de tous les classificateurs est calculée par rapport à la bonne attribution de la note à la **critique entière** et non à chaque **phrase**.

La mesure de performance d'attribution de la note à la critique entière dans le cas de deux classificateurs (statistique et linguistique) peut sembler moins précise que la mesure de performance par rapport à l'attribution de la note à chaque phrase. En effet, nous effectuons la classification de chaque phrase et non pas la moyenne des notes de toutes les phrases de chaque critique cinématographique. Donc pour ces deux classificateurs nous avons aussi effectué la mesure de la performance par rapport à la note attribuée à chaque phrase. Les résultats que nous avons obtenus n'étaient pas trop éloignés de ceux que nous avons obtenus en regardant la critique entière, pourtant les sens de la précision et du rappel sont différents dans les deux mesures. Cette validation ne peut évidemment pas être effectuée avec le classificateur de comportement des groupes. Pour cette raison nous estimons que pour pouvoir comparer les résultats de toutes les classifications nous devons tenir compte de la note attribuée à la critique entière.

La mesure de la performance d'attribution de la note par rapport à **chaque phrase** (le classificateur linguistique et statistique) demande le calcul de la précision et du rappel. Pour ces calculs nous avons besoin d'avoir :

- l'ensemble de tous les documents pertinents trouvés,
- l'ensemble de tous les documents trouvés,
- l'ensemble de tous les documents pertinents présents dans la base.

Pour l'ensemble de tous les documents pertinents trouvés, nous prenons toutes les phrases subjectives qui ont eu une note associée par le classificateur égale à la note (existante dans la base de données : note d'utilisateur) de la critique de ces phrases.
Pour l'ensemble de tous les documents trouvés, nous prenons toutes les phrases auxquelles nous avons attribué une note.
Pour l'ensemble de tous les documents pertinents présents dans la base, nous prenons toutes les phrases subjectives de la critique.

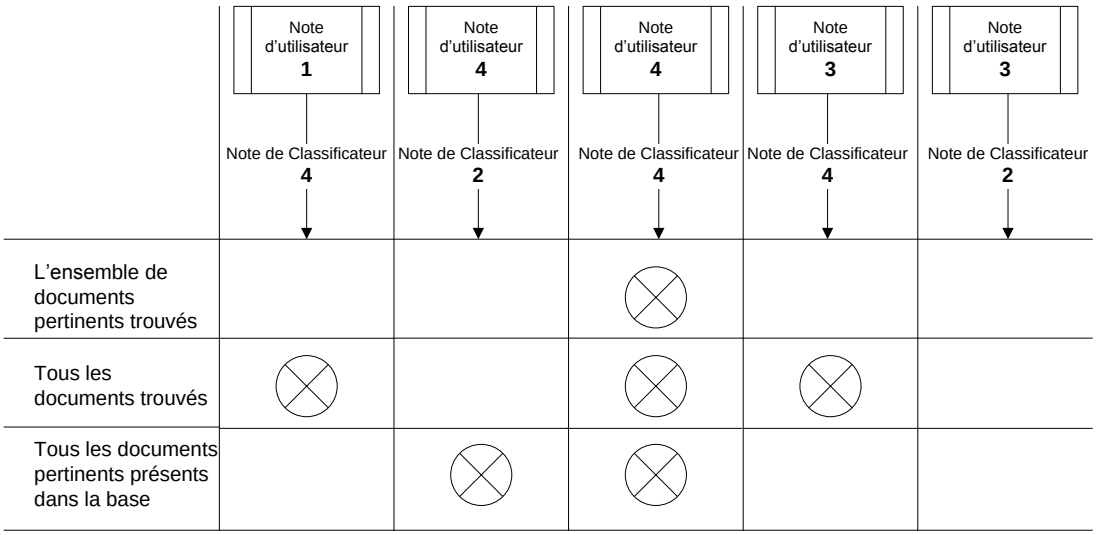

Figure 7.1: La mesure de performance d'attribution de la note par rapport à la critique entière - L'exemple montre le calcul de la précision et du rappel pour une note égale à 4

Dans le cas du calcul de la mesure de la performance d'attribution de la note par rapport à la **critique entière** pour l'ensemble de tous les documents pertinents trouvés, nous prenons toutes les critiques pour lesquelles la notation est correcte.
Pour l'ensemble de tous les documents trouvés, nous prenons toutes les critiques qui ont une note associée par la classification égale à la note pour laquelle nous avons effectué la mesure de performance.
Pour l'ensemble de tous les documents pertinents présents dans la base, nous prenons toutes les critiques qui ont une note associée par l'utilisateur égale à la note pour laquelle nous avons effectué la mesure de performance.

L'exemple de la mesure de performance par rapport à l'attribution de la note à chaque phrase est montré sur la [Figure 7.1].

Dans l'exemple présenté nous avons l'ensemble de documents pertinents trouvés dont la note est égale à 1, l'ensemble de tous les documents trouvés égale à 3 et l'ensemble de tous les documents pertinents présents dans la base égale à 2. Dans l'exemple la précision $\pi = \frac{1}{3} = 33.3\%$ et le rappel $\rho = \frac{1}{2} = 50\%$.

La base de test est constituée de 300 critiques - 60 par note. La base de test que nous avons utilisée pour calculer la performance contient :

- 828 phrases pour la note égale à 5,
- 588 phrases pour la note égale à 4,
- 657 phrases pour la note égale à 3,
- 431 phrases pour la note égale à 2,
- 1130 phrases pour la note égale à 1.

7.1.2 Le classificateur linguistique

Le classificateur linguistique utilise la base d'apprentissage pour la création des règles des grammaires locales pour chaque classe de notes. Pour effectuer la notation nous prenons une nouvelle critique de la base de test. L'attribution de la note est effectuée phrase par phrase. A la fin de processus nous obtenons un nombre des phrases avec les notes associées.

Notre base de tests contient 706 phrases objectives et 2898 phrases subjectives (744 phrases avec une note égale à 5, 533 avec une note égale à 4, 588 avec une note égale à 3, 381 avec une note égale à 2 et 893 phrases avec une note égale à 1).

	5(646)	4(458)	3(557)	2(426)	1(809)	PNC
5*(744)	539	24	44	33	29	75
4*(533)	43	377	25	27	21	40
3*(588)	15	18	399	46	22	88
2*(381)	13	12	23	238	38	57
1*(893)	12	7	39	62	681	92
PO	24	20	27	39	18	-
Précision	83.4%	82.4%	71.6%	55.9%	84.2%	-
Rappel	72.4%	70.8%	67.8%	62.5%	76.3%	-
F-score	76.5%	76.1%	69.6%	59%	80.1%	-

Tableau 7.1: Mesure de performance pour le classificateur linguistique par rapport aux phrases - en haut : la classification des phrases de chaque groupe de notation (lignes - notes des auteurs, colonnes - notes de classification), en bas : les mesures de performance

Dans le Tableau 7.1 nous montrons les résultats pour les tests du classificateur linguistique effectués sur la base de test de 300 critiques cinématographiques par la méthode de validation par le test. La mesure de performance est effectuée pour chaque phrase.
La partie haute du tableau montre la classification des phrases pour chaque groupe de note. Les colonnes répresentent les notes attribuées par notre classificateur. Les lignes représentent les critiques notées par les auteurs (Exemple : 5*(744) - correspond à 744 phrases avec une note égale à 5 selon la base de test). Les colonnes répresentent les notes attribuées par notre classificateur, les valeurs dans le tableau donnent, en détail, la répartition des notes de notre classificateur par rapport aux notes des auteurs. Dans le tableau, PO désigne les phrases objectives, PNC désigne les phrases non classées.

Dans la première colonne par exemple, 5(646) correspond à 646 phrases avec une note égale à 5 selon la note de notre classificateur, où 539 phrases correspondent à des phrases classifiées correctement, 43 correspondent à des phrases classifiées avec une note égale à 5 au lieu de 4, et ainsi de suite.
Le classificateur a attribué aussi les notes pour les phrases objectives (24 phrases pour le groupe 5, 20 pour le groupe 4, 27 pour le groupe 3, 39 pour le groupe 2 et 18 phrases pour la groupe 1). Plusieurs phrases n'ont pas été notées (75 phrases pour la note de 5, 40 pour la note de 4, 88 pour la note de 3, 57 pour la note de 2 et 92 pour la note de 1). La partie basse du tableau montre les valeurs de la précision, du rappel et du f-score pour le classificateur linguistique.

Pour calculer la note de la critique entière nous calculons la moyenne des notes de toutes les phrases notées. Nous pondérons les grammaires en fonction du niveau de l'analyse linguistique de la critique présentée dans la [*Section 6.4*]. La création des grammaires locales était effectuée en ajoutant un niveau de complexité par rapport à l'analyse linguistique. Les grammaires de niveau supérieur sont plus précises, mais le rappel est très faible. La recherche est effectuée de façon à ce qu'une phrase de la critique corresponde à une grammaire d'un niveau supérieur. Les autres grammaires de même note ne sont plus appliquées pour cette phrase. Pour cette raison nous avons la certitude que les résultats de la notation obtenus avec une telle grammaire sont plus précis.

Les grammaires ainsi que leurs pondérations ont été créées manuellement. Nous avons partagé les critiques en 4 groupes en fonction de leur niveau d'analyse linguistique. Nous avons ajouté les pondérations pour chaque groupe. Des grammaires les plus précises jusqu'aux grammaires générales les poids sont respectivement de 2.0 ; 1.6 ; 1.3 ; 1. Les poids ont été choisis pour que la valeur du F-score soit la plus performante, de manière empirique.

Dans le Tableau 7.2 nous montrons les résultats du classificateur linguistique appliqué à la base de test de 300 critiques cinématographiques par la méthode de validation par le test. La mesure de performance est effectuée pour la critique entière.

	5(60)	4(61)	3(58)	2(55)	1(66)
5*(60)	51	4	3	1	1
4*(60)	6	47	4	2	1
3*(60)	1	6	43	7	3
2*(60)	1	3	4	40	12
1*(60)	1	1	4	5	49
Précision	85%	77%	74.1%	72.7%	74.2%
Rappel	85%	78.3%	71.7%	66.7%	81.7%
F-score	85%	77.6%	72.9%	69.6%	77.8%

Tableau 7.2: Mesure de performance pour le classificateur linguistique par rapport à la critique entière - en haut : la classification des critiques de chaque groupe de notation (lignes - notes des auteurs, colonnes - notes de classification), en bas : les mesures de performance

Les résultats obtenus pour la mesure de performance par rapport à critique entière sont meilleurs que dans le cas de la mesure de performance par phrases. La raison en est que dans ce cas nous prenons la moyenne de toutes les phrases notées de la critique, les erreurs de la notation peuvent donc dans plusieurs cas être insignifiantes.

Le principal avantage de ce classificateur linguistique est qu'il donne de meilleurs résultats que les trois autres classificateurs. Il a cependant de nombreux inconvénients. Le plus important de ces inconvénients est que la réutilisation de ce classificateur dans un autre domaine demande la création de nouvelles règles de grammaire. La création de ces règles est effectuée manuellement et demande donc un temps important d'analyse et de test. Un autre problème important est qu'il est difficile de justifier mathématiquement que la forme des règles développées est la plus fiable. Autrement dit nous ne pouvons pas prouver que la forme, le nombre, et la complexité linguistique de nos grammaire sont les plus performants. Ces paramètres dans notre recherche ont été choisis empiriquement par des nombreux tests effectués à chaque étape du travail.

7.1.3 Le classificateur statistique

La base d'apprentissage est composée de 1000 critiques qui correspondent à 200 critiques par groupe de notation. La base est composée de 9289 phrases : 2264 phrases pour la note égale à 5, 1957 phrases pour la note égale à 4, 1308 pour la note égale à 3, 1925 pour la note égale à 2 et 1835 pour la note égale à 1.
Pour la représentation vectorielle nous avons calculé l'index complet qui est égal à 18422 mots lemmatisés. Pour la détection de l'opinion nous utilisons deux classificateurs de Bayes - le premier pour la détection de phrases subjectives et le deuxième pour la détection de l'intensité de l'opinion. Pour chaque classificateur nous effectuons la réduction de l'index complet et nous obtenons un ensemble de 705 mots pour le classificateur de subjectivité et un ensemble de 743 mots pour le classificateur d'intensité.

	Précision	Rappel	F-score
Classe - 5	67.5%	71.4%	69.4%
Classe - 4	71.8%	67.2%	69.4%
Classe - 3	64.2%	63.3%	63.7%
Classe - 2	63.4%	62.4%	62.9%
Classe - 1	69.3%	72.9%	71.1%

Tableau 7.3: Mesure de performance pour le classificateur statistique par rapport aux phrases

Nous utilisons la même base de test pour tous les classificateurs de notation de l'opinion. Le classificateur de subjectivité classe correctement 82,4% des phrases. Dans le Tableau 7.3 nous montrons les résultats du classificateur de l'intensité de l'opinion par la méthode de validation par test. La classification est effectuée phrase par phrase.

Comme nous l'avons déjà précisé, nous voulions comparer entre eux les résultats obtenus par chaque classificateur de la notation de l'opinion. Pour cette raison nous sommes obligés de calculer la performance par rapport à la notation de la critique entière. Nous avons procédé de la même manière que pour le classificateur linguistique sauf que dans ce cas nous n'avons attribué aucune pondération aux phrases. Nous avons attribué sa note à la critique en calculant la note moyenne de toutes les phrases. Dans

	5(63)	4(56)	3(58)	2(57)	1(66)
5*(60)	43	10	2	4	1
4*(60)	12	41	4	3	0
3*(60)	3	3	39	8	7
2*(60)	3	2	6	37	12
1*(60)	2	0	7	5	46
Précision	68.3%	73.2%	67.2%	64.9%	69.7%
Rappel	71.7%	68.3%	65%	61.7%	76.7%
F-score	70%	70.7%	66.1%	63.3%	73%

Tableau 7.4: Mesure de performance pour le classificateur statistique par rapport à la critique entière - en haut : la classification des critiques de chaque groupe de notation (lignes - notes des auteurs, colonnes - notes de classification), en bas : les mesures de performance

le Tableau 7.4 nous montrons les résultats pour le classificateur statistique obtenus sur la base de test de 300 critiques cinématographiques par la méthode de validation par test.

7.1.4 Classification des sentiments par phrases

Nous avons présenté les résultats de la mesure de performance pour deux classificateurs, l'un linguistique et l'autre statistique. Ces sont les classificateurs qui traitent la critique phrase par phrase. Nous montrons la comparaison de ces deux approches en montrant la valeur de la précision [Figure 7.2], du rappel [Figure 7.3] et du F-score [Figure 7.4] pour chaque groupe de critique.

Nous pouvons constater que le classificateur linguistique donne de meilleurs résultats que le classificateur statistique. Nous avons donc réussi à appliquer l'analyse linguistique pour la mesure de l'intensité des sentiments.

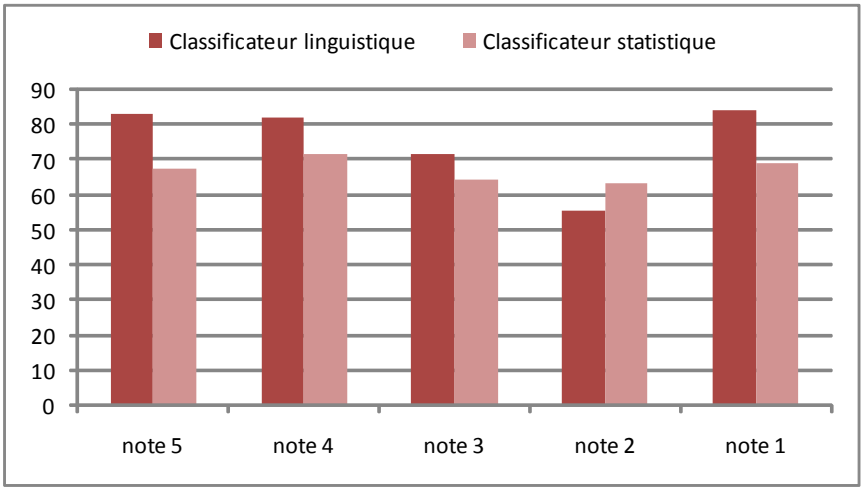

FIGURE 7.2: Précision pour la classification par phrases - Comparaison des résultats des classificateurs (linguistique et statistique) en mesurant la précision (classification par phrase)

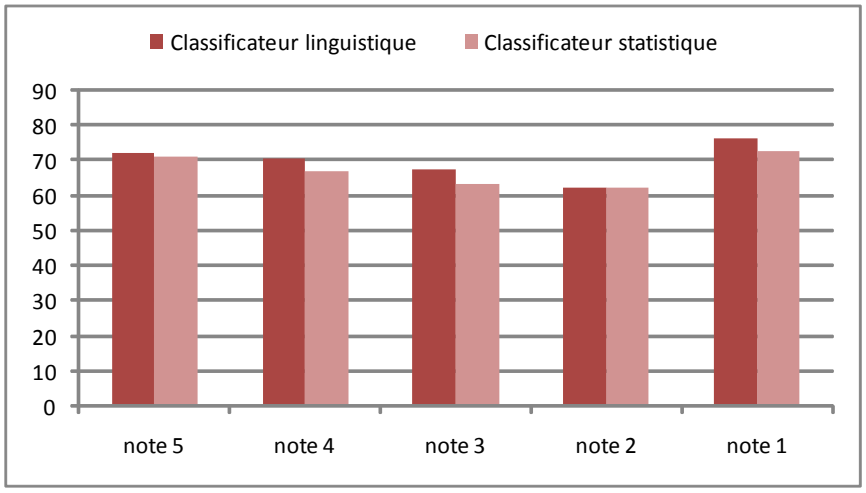

FIGURE 7.3: Rappel pour la classification par phrases - Comparaison des résultats des classificateurs (linguistique et statistique) en mesurant le rappel (classification par phrase)

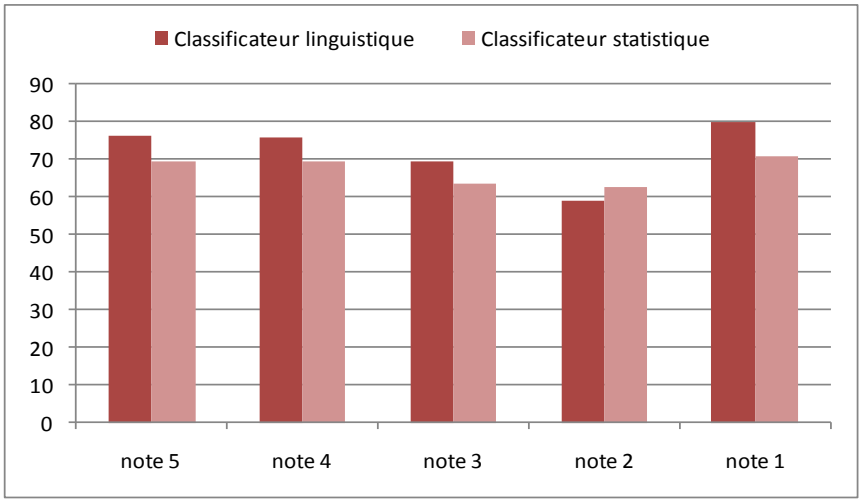

Figure 7.4: F-score pour la classification par phrases - Comparaison des résultats des classificateurs (linguistique et statistique) en mesurant la F-score (classification par phrase)

7.1.5 Le classificateur de comportement des groupes

Pour le classificateur de comportement des groupes nous avons utilisé la base d'apprentissage pour déterminer le comportement de chaque groupe composé par les critiques qui ont la même note associée. Pour effectuer le processus de la notation nous prenons une nouvelle critique de la base de test. La détermination du comportement globale de chaque groupe permet de déterminer à quel groupe appartient une nouvelle critique cinématographique. Pour les nouveaux critiques nous calculons la distance euclidienne entre ses caractéristiques et les caractéristiques des groupes. La note attribuée à la critique est celle pour laquelle la distance est la plus courte. Dans le Tableau 7.5 nous montrons les résultats du classificateur de comportement des groupes appliqué à la base de test de 300 critiques cinématographiques par la méthode de validation par le test. La mesure de performance est effectuée par la critique entière.

Comme nous pouvons le constater les résultats obtenus par cette classification sont moins fiables que ceux obtenus par la classification linguistique. Pourtant il faut préciser que les résultats sont légèrement meilleurs que ceux obtenus par la classification de Bayes. Un grand avantage de cette classification (contrairement au classificateur linguistique) est la facilité de sa réutilisation dans un nouveau domaine, celle-ci ne demandant pas beaucoup de travaille manuel. Il suffit d'appliquer une nouvelle base d'apprentis-

	5(62)	4(61)	3(56)	2(58)	1(63)
5*(60)	45	6	3	2	4
4*(60)	11	43	2	3	1
3*(60)	3	7	41	6	3
2*(60)	2	3	6	39	13
1*(60)	1	2	3	8	42
Précision	72.6%	70.5%	73.2%	67.2%	66.7%
Rappel	75%	71.6%	68.3%	65%	70%
F-score	73.8%	71%	70.8%	66.1%	68.3%

Tableau 7.5: Mesures de performance pour le classificateur de comportement des groupes - en haut : la classification des critiques de chaque groupe de notation (lignes - notes des auteurs, colonnes - notes de classification), en bas : les mesures de performance

sage pour pouvoir calculer les nouvelles caractéristiques des groupes. Le travail manuel nécessaire consiste uniquement à rechercher les mots et les expressions caractéristiques de ce nouveau domaine (qui devraient être peu éloignés de ceux présentés) et de fournir de nouveaux critères pour la recherche. Un des principaux inconvénients de cette classification est qu'elle nécessite une très grande base d'apprentissage pour pouvoir rechercher les caractéristiques de comportement des groupes. Cela n'est pas génant dans notre cas, car nous avons une très grande base de critiques cinématographiques déjà notées. L'utilisation de cette méthode dans un domaine où l'on ne dispose pas de ces ressources est remise en question car l'annotation manuelle des données demanderait beaucoup trop de temps.

7.1.6 Classification des sentiments par la critique entière

Nous avons présenté les résultats de la mesure de performance pour les trois classificateurs : linguistique, statistique et de comportement des groupes. Nous avons présenté la comparaison de toutes les approches appliquées à la notation des sentiments. Nous avons mis en valeur la comparaison de la précision [Figure 7.5], du rappel [Figure 7.6] et du F-scores [Figure 7.7] pour chaque groupe de critique.

Nous pouvons constater encore une fois que le classificateur linguistique donne de meilleurs résultats. Les résultats obtenus grâce au classificateur de comportement des groupes sont légèrement meilleurs que ceux obtenus par le classificateur statistique.

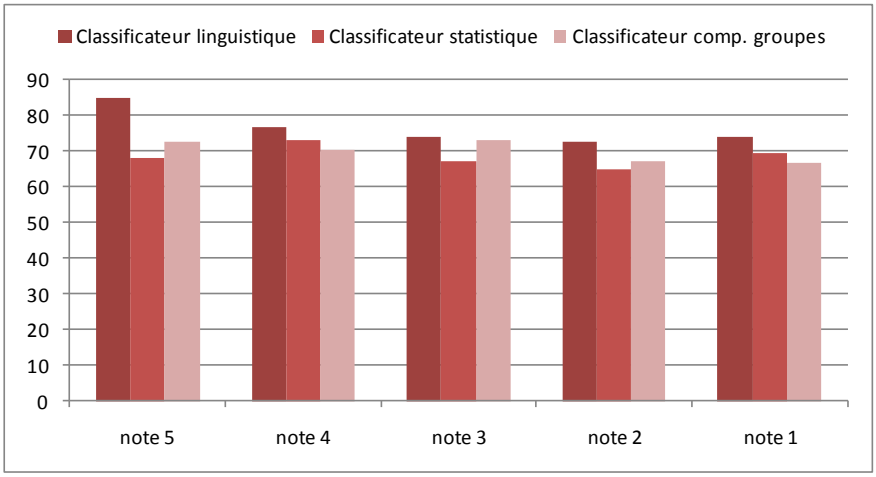

Figure 7.5: Précision pour la classification par la critique entière - Comparaison des résultats de tous les classificateurs de la notation de l'opinion en mesurant la précision

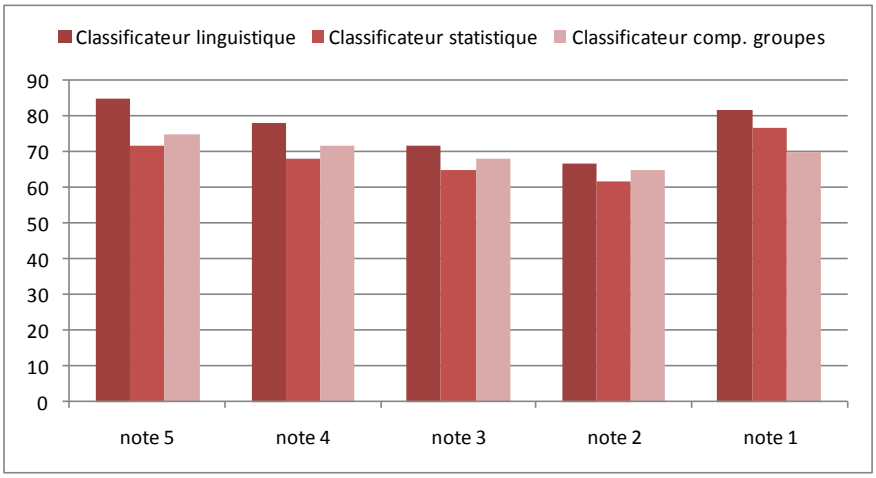

Figure 7.6: Rappel pour la classification par la critique entière - Comparaison des résultats de tous les classificateurs de la notation de l'opinion en mesurant le rappel

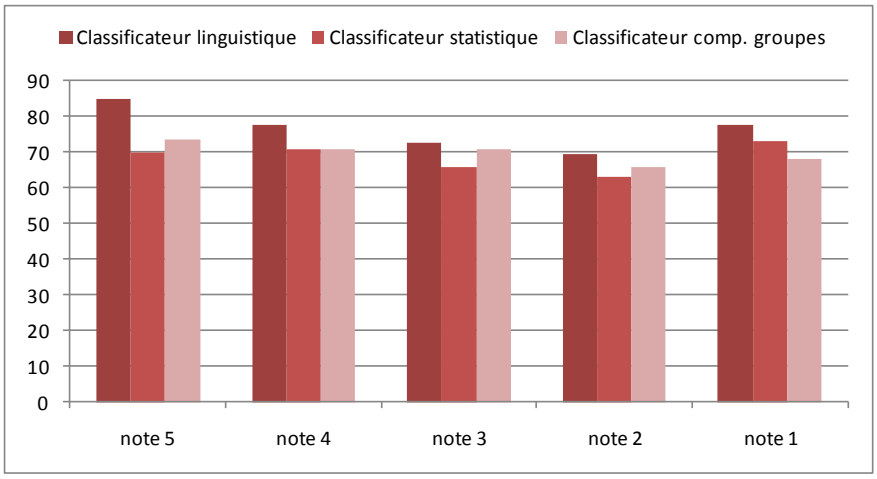

Figure 7.7: F-score pour la classification par la critique entière - Comparaison des résultats de tous les classificateurs de la notation de l'opinion en mesurant le F-score

7.2 Les tests de classification finale

Comme nous l'avons remarqué nous avons un classificateur déterministe dans plusieurs situations. Nous avons donc amélioré nos résultats obtenus en utilisant les réseaux de neurones. Pour cette étape de classification nous avons fondé notre approche uniquement sur les résultats des 3 classificateurs décrits précédemment. Les résultats finaux obtenus par le calcul de la moyenne basée uniquement sur les notes entières de chaque classificateur (1 à 5) sont moins bonnes (la précision et le rappel) que les résultats obtenus par le meilleur classificateur - classificateur linguistique. Pourtant nous avons amélioré nos résultats par l'utilisation de réseaux de neurones en prenant en considération chaque probabilité de chaque note de chaque classificateur. Ces résultats ont été améliorés d'un ordre de 4% par rapport au résultat du meilleur classificateur - le classificateur linguistique [Figure 7.8].

Le F-score calculé par rapport à la note finale est de 83,1% pour 5*, 81,2% pour 4*, 74,5% pour 3*, 72,2% pour 2* et 81,4% pour 1*. Pour l'apprentissage de ce classificateur nous avons utilisé une nouvelle base d'apprentissage et la même base de tests que pour les trois classificateurs présentés précédemment.

L'utilisation des réseaux de neurones est justifiée par la présence d'une très grande base d'apprentissage - la base des critiques déjà notées. Un avantage de ce classificateur

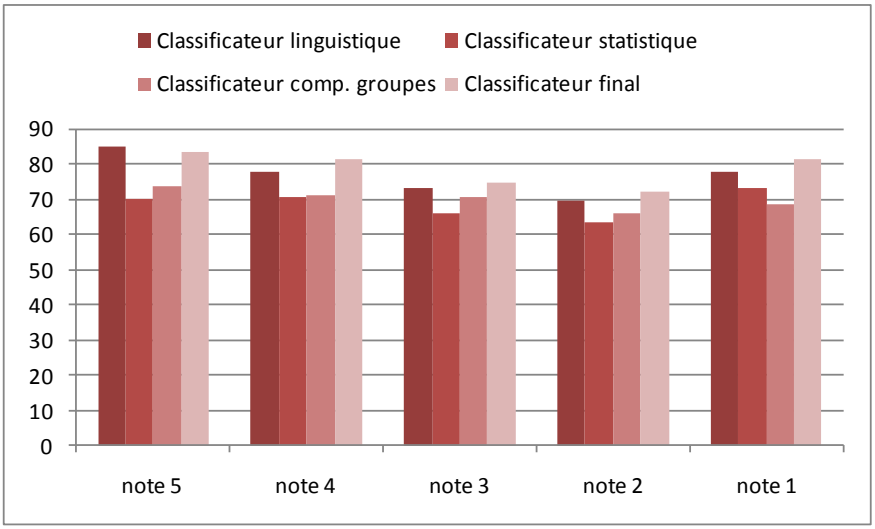

Figure 7.8: Les résultats du classificateur final - Comparaison des résultats de toutes les classifications

est qu'il ne demande pas de prétraitement spécial de données. Le classificateur final n'est pas utilisé pour la notation des sentiments. Cependant il est utilisé pour combiner les résultats obtenus des trois classificateurs présentés dans cette thèse.

7.3 Conclusion

Nous avons remarqué que nous obtenons de meilleurs résultats avec le classificateur linguistique (surtout la précision). Les moins bons résultats sont ceux du classificateur statistique de "naïf Bayes" (le rappel est correct mais la précision est faible). Cela démontre la nécessité d'une analyse linguistique profonde. Nous avons observé que les meilleurs résultats ont été obtenus dans chaque approche pour des opinions extrêmes. Il est plus facile de noter automatiquement et de juger les critiques cinématographiques ayant des notes de 1 ou 5. Cela semble évident, car les émotions extrêmes sont plus fortes et généralement la personne les exprime de manière plus intense. De plus le texte des opinions extrêmes est plus long ce qui favorise l'attribution d'une note correcte. Partant du principe qu'il est nécessaire de disposer de grammaires plus complexes, nous avons montré que le classificateur linguistique donne de meilleurs résultats que le classificateur statistique ou le classificateur de comportement de groupe.

Un point important à la vue des résultats obtenus est la réussite de l'implémentation de l'approche linguistique, ce qui démontre l'importance de l'utilisation de l'analyse linguistique dans le domaine de l'*Opinon Mining*.