۳-۷ روش‌های وزن دهی ویژگی

ویژگی : کلماتی که به مفاهیم اصلی متن نزدیک‌تر بوده و دربردارنده مهمترین موضوعات یک مستند هستند.

روش‌های متنوعی برای وزن دهی ویژگی‌ها گزارش شده است. از میان این روشها می‌توان به روش‌های مبتنی بر تعداد تکرار کلمه[۱] (TF)، روش‌های مبتنی بر تعداد تکرار کلمه در مستندات مختلف[۲] (IDF)، روش‌های ترکیبی TF و IDF، روش‌های مبتنی بر الگوریتم ژنتیک و شبکه‌های عصبی، روش‌های مبتنی بر انتخاب ویژگی، روش‌های مبتنی بر اطلاعات طبقات اشاره کرد .

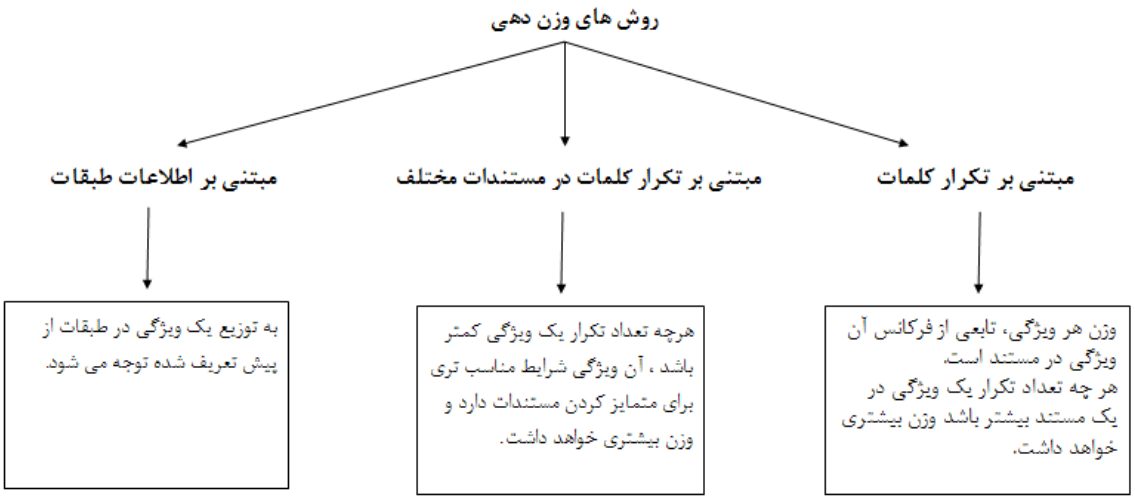
محتوای هر مستند $d_i \in D$ با برداری در فضای ویژگی‌ها $d_i = (w_{1i}, ..., w_{ki})$ نشان داده می‌شود که در آن k تعداد ویژگی‌های متمایز در کل مجموعه D و w_{ki} وزن ویژگی t_k است که نشان می‌دهد که ویژگی t_k تا چه حد می‌تواند معنای مستند d_i را در بر داشته باشد.

به طور کلی می‌توان روش‌های وزن دهی ویژگی که در حوزه طبقه بندی مستندات استفاده می‌شوند را به روش‌های مبتنی بر TF [۳] ، روش‌های مبتنی بر IDF [۴] و در نهایت روش‌های مبتنی بر اطلاعات طبقات دسته بندی نمود که در این قسمت خلاصه ای از نحوه عملکرد هر کدام ذکر می‌گردد.

بدیهی است روش‌های وزن دهی ویژگی شرح داده شده در این قسمت شامل کلیه روش‌های وزن دهی ویژگی مطرح شده در حوزه طبقه بندی مستندات نبوده و تنها به معرفی روش‌های شناخته شده در این حوزه می‌پردازیم.

---

[1] Term Frequency
[2] Inverse Document Frequency
[3] Term Frequency
[4] Inverse Documeny Frequency

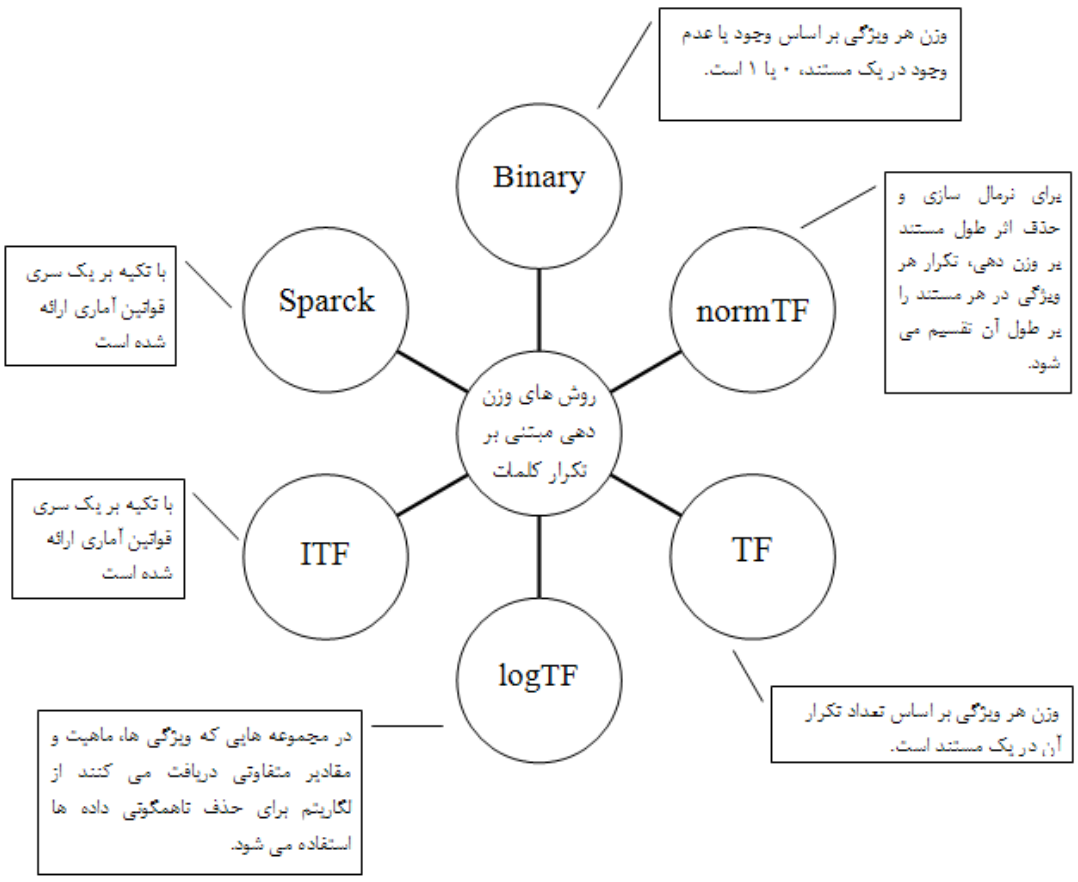


شکل ۴-۳: روش های وزن دهی (مرحله دوم طبقه بندی)

۳-۷-۱ روش‌های وزن دهی ویژگی مبتنی بر TF

در این روش‌ها، وزن‌دهی ویژگی‌ها تابعی از توزیع ویژگی‌های مختلف در هر یک از مستندات $d_i \in D$ می‌باشد.

به بیان دیگر در این دسته از روش‌ها، وزن هر ویژگی، تابعی از فرکانس آن ویژگی در مستند است. ایده اصلی وزن‌دهی ویژگی‌ها در این دسته این است که هر چه تعداد تکرار یک ویژگی در یک مستند بیشتر باشد آن ویژگی قابلیت بیشتری در نشان دادن معنای مستند و متمایز کردن آن مستند از سایر مستندات را دارا بوده و بایستی وزن بیشتری به خود اختصاص دهد.



شکل ۴-۴: روش های وزن دهی مبتنی بر تکرار کلمات در یک مستند

الف – روش وزن دهی باینری

یکی از ساده ترین روش‌های موجود برای وزن‌دهی ویژگی روش وزن‌دهی باینری (دودویی) است که در آن وزن هر ویژگی t_k در مستند d_i بر اساس وجود یا عدم وجود آن ویژگی در مستند مربوطه برابر ۱ یا صفر خواهد بود.

در این روش فرکانس ویژگی یعنی تعداد تکرار ویژگی در مستند اصلاً در نظر گرفته نمی شود و فقط وجود یا عدم وجود ویژگی در مستند تعیین کننده وزن آن خواهد بود.

$$w_{ki} = tf(t_k, d_i) = \begin{cases} 1 & t_k \in \text{vector of } d_i \\ 0 & t_k \notin \text{vector of } d_i \end{cases} \qquad (۶-۳)$$

این روش برای الگوریتم‌های طبقه بندی مبتنی بر یادگیری ماشینی نظیر بیزین و درخت‌های تصمیم‌گیری کاربرد دارد.

ب – روش وزن دهی TF

این روش ساده و بسیار کاربردی مشابه روش باینری است با این تفاوت که در صورت وجود ویژگی t_k در مستند d_i وزن آن برابر تعداد تکرار آن ویژگی در مستند مربوطه می‌باشد. رابطه (۳-۷) روش محاسبه وزن ویژگی t_k در مستند d_i را نشان می‌دهد.

$$w_{ki} = tf(t_k, d_i) = \begin{cases} \#(t_k, d_i) & t_k \in \text{vector of } d_i \\ 0 & t_k \notin \text{vector of } d_i \end{cases} \qquad (۳-۷)$$

که در آن $\#(t_k , d_i)$ برابر تعداد تکرار هر ویژگی t_k در مستند d_i است.

پ – روش normTF

به طور معمول طول مستندات مجموعه مستندات D برابر نمی باشد. این امر در حوزه بازیابی مستندات باعث می‌شود تا مستنداتی که دارای طول بیشتر (تعداد کلمات بیشتر) هستند شانس بالاتری جهت بازیابی داشته باشند. در حوزه طبقه بندی مستندات نیز هرچه طول مستند بیشتر باشد تعداد تکرار ویژگی‌ها در آن بیشتر می‌شود در نتیجه ویژگی‌های موجود در مستندات طولانی تر وزن بیشتری به خود اختصاص می‌دهند که اصلاً معیار مناسبی جهت وزن دهی نیست.

لذا در این روش برای حذف اثر طول مستند بر روی نحوه وزن دهی ویژگی‌های آن و محدود کردن مقدار وزن ویژگی‌ها بین محدوده $(0,1)$ از نرمال سازی استفاده می‌شود. برای نرمال سازی تنها کافی است فرکانس تکرار هر ویژگی در هر مستند را بر طول آن مستند تقسیم نمود.

با این کار اثر نا مطلوب طول مستند بر وزن دهی ویژگی از بین می‌رود. در رابطه (۳-۸) روش محاسبه وزن ویژگی t_k در مستند d_i نشان داده شده است.

$$w_{ki} = normTF(t_k, d_i) = \frac{tf(t_k, d_i)}{\sqrt{\sum_k (tf(t_k, d_i))^2}} \qquad (۳-۸)$$

$tf(t_k, d_i)$ از رابطه مربوط به روش TF بدست می‌آید و برابر تعداد تکرار هر ویژگی t_k در مستند d_i است.

ت – روش logTF

در برخی مجموعه‌های داده‌ای متفاوت بودن ماهیت ویژگی‌ها و مقادیری که به خود اختصاص می‌دهند می‌تواند بر روی دقت و کارایی الگوریتم طبقه بندی کننده تاثیر منفی بگذارد.

به طور مثال مقداری که به ویژگی سن فرد داده می‌شود با مقداری که به ویژگی حقوق فرد داده می‌شود کاملا با یکدیگر متفاوت بوده و نباید این تفاوت باعث شود که ویژگی حقوق فرد فقط به صرف بالا بودن مقداری که دارد وزن بیشتری به خود اختصاص دهد.

به همین جهت از عملگر لگاریتم برای حذف این اثر نامطلوب و یکسان کردن محدوده مقادیر تخصیصی به هریک از ویژگی‌ها استفاده می‌شود:

$$W_{KI} = \log TF(t_k, d_i) = \log(tf(t_k, d_i)) \qquad (۹-۳)$$

که $tf(t_k, d_i)$ از رابطه مربوط به روش TF به دست آمده است.

ث – روش ITF

روش وزن دهی ویژگی ITF برای اولین بار توسط آقای Leoplod در [Leopold] ارائه شد که بر اساس آن وزن هر ویژگی از رابطه (۳-۱۰) محاسبه می‌شود:

$$w_{ki} = ITF(t_k, d_i) = 1 - \frac{r}{r + tf(t_k, d_i)} \qquad (۱۰-۳)$$

که معمولاً مقدار r برابر ۱ قرار داده می‌شود. این رابطه با تکیه بر یک سری قوانین آماری ارائه شده است.

ج – روش Sparck

این روش که اولین بار توسط Sparck ارائه شد. از تئوری‌های آماری برای وزن دهی به ویژگی‌ها بهره برده است.

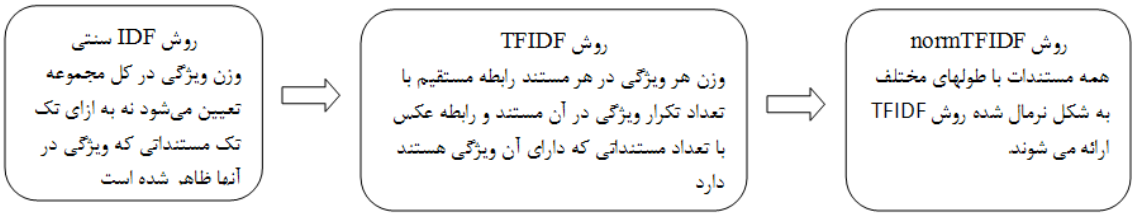
$$w_{ki} = \text{Sparck}(t_k , d_i) = tf(t_k , d_i) * (k - \log(p_k)) \qquad (۱۱-۳)$$

که در آن k تعداد کل ویژگی‌های متمایز در مجموعه $P_K = \Sigma tf(t_k , d_i)$ می‌باشد.

این روش وزن دهی ویژگی، در حوزه بازیابی اطلاعات مطرح شده و تاکنون در حوزه طبقه بندی مستندات استفاده نشده است.

۳ – ۷ – ۲ روش‌های وزن دهی ویژگی مبتنی بر IDF

در این روش‌ها، وزن دهی ویژگی‌ها تابعی از توزیع ویژگی t_k در داخل مجموعه مستندات D است. ایده اصلی وزن دهی در این دسته به این صورت است که هر چه تعداد مستنداتی که دارای ویژگی t_k هستند کمتر باشد، t_k ویژگی مناسب تری برای متمایز کردن مستندات از یکدیگر بوده و بایستی وزن بیشتری به خود اختصاص دهد.



شکل ۴-۵ : روش های وزن دهی مبتنی بر تکرار کلمات در مستندات مختلف

الف – روش IDF سنتی

این روش که اولین بار در حوزهٔ بازیابی اطلاعات مطرح شده است به شکل زیر است:

$$w_{ki} = idf(t_k) = \log \frac{|D|}{|D(t_k)|} \qquad (۱۲-۳)$$

که در آن:

|D| : تعداد کل مستندات مجموعه D

$|D(t_k)|$: تعداد مستندانی از مجموعه D که ویژگی t_k در آنها وجود دارد.

بدیهی است که w_{ki} در رابطه مذکور با افزایش $|D(t_k)|$ کاهش می‌یابد.

ملاحظه می‌شود که در این روش وزن ویژگی در کل مجموعه تعیین می‌شود نه به ازای تک تک مستنداتی که ویژگی در آنها ظاهر شده است.

ب – روش TFIDF

روش وزن دهی ویژگی TFIDF که از **رایج ترین** روش‌های وزن دهی ویژگی است اولین بار در حوزهٔ بازیابی اطلاعات مطرح و سپس در طبقه بندی مستندات برای وزن دهی ویژگی‌ها از آنها استفاده شده است. در این روش از ترکیب دو روش TF و IDF به صورت زیر استفاده شده است.

$$w_{ki} = tfidf(t_k, d_i) = tf(t_k, d_i) * idf(t_k) = tf(t_k, d_i) * \log \frac{|D|}{|D(t_k)|} \qquad (۱۳-۳)$$

همانطور که مشخص است در این روش، وزن هر ویژگی در هر مستند رابطه مستقیم با تعداد تکرار ویژگی در آن مستند و رابطه معکوس با تعداد مستنداتی که دارای آن ویژگی هستند دارد.

پ – روش normTFIDF

برای اطمینان از اینکه همه مستندات با طولهای مختلف شانس برابری برای بازیابی شدن داشته باشند روش TFIDF فوق به صورت نرمال به شکل زیر ارائه شده است:

$$w_{ki} = normTFIDF(t_k, d_i) = \frac{tfidf(t_k, d_i)}{\sqrt{\sum_k (tfidf(t_k, d_i))^2}} \qquad (۱۴-۳)$$

۳ – ۷ – ۳ روش‌های وزن دهی ویژگی مبتنی بر اطلاعات طبقات [1]

روش TFRF

[1] Class-Based Feature Weighting Method

در این روش تنها به توزیع ویژگی t_k در مستند d_i و یا توزیع ویژگی t_k در مجموعه D بسنده نکرده و از توزیع ویژگی t_k در طبقات از پیش تعریف شده $c_i \in C$ استفاده می‌شود. که مورد نیاز در بحث ما نیست.

معیارهای سنجش

در این بخش از شاخص‌های زیر برای ارزیابی عمکرد هر یک از روش‌های وزن‌دهی استفاده می‌شود. این شاخص‌ها عبارتند از: معیار دقت (AC)، معیار صحت (Pr)، میزان پوشش (Re)، معیار F، میانگین میکرو و میانگین ماکرو .

قبل از پرداختن به هر یک از معیارها لازم است چند اصطلاح معرفی گردد:

TP: تعداد مواردی که درست تشخیص داده شده است (مثلا در شناسایی خودکار کلیدواژه‌ها). پس عملکرد در این زمینه درست بوده است.

FP: تعداد مواردی که شناسایی صحیح نبوده است. پس عملکرد در این زمینه اشتباه بوده است.

FN: تعداد مواردی که شناسایی نشده اند اما باید شناسایی می‌شدند.

TN: تعداد مواردی که شناسایی نشده اند و واقعا هم نباید تشخیص داده می‌شدند. پس عملکرد در این زمینه درست بوده است.

	مربوط	نامربوط
بازیابی شده	مثبت واقعی TP	مثبت کاذب FP
بازیابی نشده	منفی کاذب FN	منفی واقعی TN

بر اساس هر یک از تعاریف بالا :

معیار دقت (AC)[1] که نشان دهنده میزان دقت کار است :

$$Ac(c_j) = \frac{TP(c_j) + TN(c_j)}{TP(c_j) + FP(c_j) + TN(c_j) + FN(c_j)} \qquad (۵-۱)$$

صورت کسر : جمع مواردی که درست انجام داده است . (چه در معرفی کردن و چه در معرفی نکردن)

مخرج کسر : کل موارد (مستند، کلمه، جمله یا هر چیزی که به عنوان منبع اصلی برای پیاده سازی الگوریتم بر اساس آنها کار صورت پذیرفته است)

معیار صحت (Pr)[2] این معیار به عنوان یکی از معروف‌ترین معیارهای ارزیابی بوده و میزان صحت عملکرد را نشان می‌دهد:

[1] Accuracy

[2] Precision

$$\Pr(c_j) = \frac{TP(c_j)}{TP(c_j) + FP(c_j)} \qquad (۵-۲)$$

صورت کسر : موارد درست یافت شده

مخرج کسر : کل موارد یافت شده (غلط و صحیح)

میزان پوشش (Re)[۱] این معیار نسبت تعداد مواردی که به طور صحیح شناسایی شده اند به تعداد کل مواردی که بایستی شناسایی می‌شدند را نشان می‌دهد

$$\mathrm{Re}(c_j) = \frac{TP(c_j)}{TP(c_j) + FN(c_j)} \qquad (۵-۳)$$

صورت کسر : موارد یافت شده درست

مخرج کسر : مواردی که باید یافت می شده است .

معیار F: از آنجایی که هیچ یک از معیارهای مذکور به تنهایی برای اندازه‌گیری کارایی ندارد و ممکن است به تنهایی نتایج نادرستی ارائه دهند لذا این معیار ترکیبی از معیار صحت و میزان پوشش می‌باشد:

$$F_1(c_j) = \frac{2 * Pr(c_j) * Re(c_j)}{Pr(c_j) + Re(c_j)} \qquad (۵-۴)$$

[۱] Recall

به عنوان مثال در یک مستند بعد از حذف کلمات توقف، ۵۰ اصطلاح استخراج شد.

۷ کلیدواژه توسط کاربر انسانی انتخاب شد. الگوریتم پیاده سازی شده برای استخراج کلیدواژه ۹ کلمه را انتخاب کرده است که از میان این ۹ اصطلاح ، ۵ عبارت با کلمات یافت شده توسط کاربر انسانی تطابق داشته و ۴ کلمه به اشتباه انتخاب شده است.

بنابراین:

TP = 5

FP = 4

FN = 2

TN = 39

AC = (5 + 39) / 50 = 0/88

Pr = 5 / (5+4) = 0/55

Re = 5 / (5+2) = 0/71

F = (2 * 0/55 * 0/71) / (0/55 + 0/71) = 0/6198

اندازه‌گیری توافق

برای بررسی دقت و کیفیت مجموعه اطلاعات جمع‌آوری شده لازم است این موارد توسط اشخاص دیگر (از این به بعد به آنها حاشیه‌نویس گفته می‌شود) نیز بررسی گردد. میزان توافق یا عدم توافق آنها بر روی اطلاعات جمع‌آوری شده به عنوان معیاری برای کیفیت اطلاعات محسوب می‌شود.

در چنین شرایطی و به منظور بررسی میزان توافق بین حاشیه‌نویسان ضریب کاپای کوهن[1] محاسبه می‌شود. از این ضریب برای ارزیابی سازگاری نظرات حاشیه‌نویسان به عنوان یک ابزار اندازه‌گیری کیفیت و دقت اطلاعات استفاده می‌شود.

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)} \quad (5\text{-}5)$$

(۵-۶)

$$P(A) = \frac{A + D}{A + B + C + D} \quad (5\text{-}7)$$

$$P(E) = \frac{(A + B) * (A + C)}{(A + B + C + D)^2} + \frac{(B + D) * (C + D)}{(A + B + C + D)^2}$$

در این فرمول P(A) نسبت سازگاری و P(E) احتمال سازگاری است.

اما برای بدست آوردن مقادیر مربوط به A، B، C و D از جدول زیر استفاده می‌شود.

تعداد توافقات		حاشیه نویس ۱	
		Yes	No
حاشیه نویس ۲	Yes	A	B
	No	C	D

بر اساس جدول بالا مجموع کل نظرات هر دو حاشیه‌نویس (A+B+C+D) است.

A : تعداد مواردی که هر دو حاشیه‌نویس نسبت به داده‌ها توافق داشته‌اند.

B : تعداد مواردی که حاشیه‌نویس ۱ مخالف و حاشیه نویس ۲ موافق بوده است.

[1] Cohen's Kappa coefficient

C : تعداد مواردی که حاشیه‌نویس ۱ موافق و حاشیه نویس ۲ مخالف بوده است.

D : تعداد مواردی که هر دو حاشیه‌نویس مخالف بوده‌اند.

آماره کاپا یک واحد اندازه‌گیری توافقی بین قضاوت در ارزیابی و در نظر گرفتن شانس توافق است. در تفسیر عدد کاپا باید توجه داشت که بیشترین مقدار ۱ است. مقادیر از ۰/۸ تا ۱ را به عنوان توافق عالی و مقادیر بین ۰/۶ تا ۰/۸ را توافق قابل قبول و ۰/۴ تا ۰/۶ را متوسط قلمداد می‌کنند. مقادیر کمتراز ۰/۴ بی ارزش است.

جدول زیر مربوط به میزان توافق حاشیه‌نویسان مربوط به مجموعه کلاس‌های معنایی فعل است. این مجموعه شامل ۱۴۹۰۳ فعل در غالب ۲۰۳۸ گروه فعلی بوده است.

تعداد توافقات		حاشیه نویس ۱	
		Yes	No
حاشیه نویس ۲	Yes	۱۲۱۷۱	۲۵۷
	No	۳۴۱	۲۴۳۴

که بر اساس محاسبات، ضریب کاپای این مجموعه برابر با ۰/۸۶ به دست آمد.

جدول زیر مربوط به میزان توافق حاشیه‌نویسان مربوط به قطبیت افعال مرکب است.

تعداد توافقات		حاشیه نویس ۱	
		Yes	No
حاشیه نویس ۲	Yes	۱۰۷۶	۱۳۴
	No	۱۶۳	۸۹۷

که بر اساس محاسبات، ضریب کاپای این مجموعه برابر با ۰/۷۳ به دست آمد.

به عنوان اندازه‌گیری دیگر برای توافق بین حاشیه‌نویسان، اندازه‌گیری F محاسبه می‌شود. اندازه گیری F به صورت زیر فرمول‌بندی می‌شود:

$$F - measure = \frac{\frac{m}{C1} * \frac{m}{C2} * 2}{\frac{m}{C1} + \frac{m}{C2}} = \frac{m * 2}{C1 + C2}$$

که در آن:

C1 و C2 به ترتیب تعداد نمونه‌های مشخص شده توسط هر حاشیه‌نویس و m تعداد نمونه‌های تطبیقی است.