

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



دانشگاه کردستان  
دانشکده مهندسی  
گروه مهندسی کامپیوتر و فناوری اطلاعات

عنوان:

بهبود طبقه‌بندی متن با استفاده از روش‌های ترکیب

پژوهشگر:  
علی دانش

اساتید راهنما:  
دکتر فردین اخلاقیان  
دکتر بهروز مینایی

پایان‌نامه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش مصنوعی

اسفند ماه ۱۳۹۱

کلیه حقوق مادی و معنوی مترتب بر نتایج مطالعات،  
ابتکارات و نوآوری‌های ناشی از تحقیق موضوع  
این پایان‌نامه متعلق به دانشگاه کردستان است.

## \*\*\* تعهد نامه \*\*\*

اینجانب علی دانش دانشجوی کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش مصنوعی دانشگاه کردستان، دانشکده مهندسی گروه کامپیوتر و فناوری اطلاعات تعهد می‌نمایم که محتوای این پایان‌نامه نتیجه تلاش و تحقیقات خود بوده و از جایی کپی‌برداری نشده و به پایان رسانیدن آن نتیجه تلاش و مطالعات مستمر اینجانب و راهنمایی و مشاوره اساتید بوده است.

با تقدیم احترام

علی دانش

۱۳۹۱/۱۲/۲۳



دانشگاه کردستان  
دانشکده مهندسی  
گروه مهندسی کامپیوتر و فناوری اطلاعات

پایان نامه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش مصنوعی

عنوان:

بهبود طبقه بندی متن با استفاده از روش های ترکیب

پژوهشگر:

علی دانش

در تاریخ ۱۳۹۱/۱۲/۲۳ توسط کمیته تخصصی و هیأت داوران زیر مورد بررسی قرار گرفت و با نمره .....  
و درجه ..... به تصویب رسید.

<u>امضاء</u>	<u>مرتبۀ علمی</u>	<u>نام و نام خانوادگی</u>	<u>هیأت داوران</u>
	استادیار	دکتر فردین اخلاقیان	۱- استاد راهنما
	استادیار	دکتر محمد رزاقی	۲- استاد داور خارجی
	استادیار	دکتر پرهام مرادی	۳- استاد داور داخلی

مهر و امضاء معاون پژوهشی و تحصیلات تکمیلی دانشکده

مهر و امضاء گروه

# سپاسگذاری و قدردانی

شکر خداوند متعال را به جای آورده که توفیق نصیب من کرد تا این پایان نامه را به سرانجام

برسانم.

در آغاز لازم می دانم از پدر، مادر و همسر به دلیل حمایت ها و تشویق هایشان که باعث رشد و

پیشرفت اینجانب گردید کمال تشکر را داشته باشم.

همچنین از تمامی اساتید و دوستانی که مرا در انجام مراحل مختلف این پایان نامه یاری نمودند،

بخصوص اساتید راهنمای گرامی جناب آقایان دکتر اخلاقیان و دکتر مینایی که در تمام مراحل انجام

این پایان نامه پشتیبان بنده بودند، تشکر ویژه ای دارم.

در پایان از جناب آقای دکتر مرادی و جناب آقای دکتر رزاقی که برای داوری این پایان نامه قبول

زحمت نمودند و وقت گران بهای خود را در اختیار اینجانب قرار دادند، تشکر می کنم.

با تقدیم احترام

علی دانش

## چکیده

با توجه به گسترش اینترنت و افزایش چشمگیر حجم مطالب متنی، ابزارها و روش‌های خودکار برای مدیریت اسناد و مطالب متنی، مورد توجه بسیار زیادی قرار گرفته‌اند. از جمله‌ی این ابزارها می‌توان به طبقه‌بند متن اشاره کرد که می‌تواند در این زمینه بسیار مؤثر باشد. این ابزار کاربردهای فراوانی در سیستم‌های بررسی متون مانند موتورهای جستجو، سیستم‌های فیلترینگ، تشخیص هرزنامه‌ها، کتابخانه‌های دیجیتال و سایر سیستم‌های مشابه دارد.

از آن جایی که صحت طبقه‌بندی متن یکی از مهمترین شاخص‌های موفقیت این سیستم‌ها به شمار می‌آید، از اهداف اساسی این پایان‌نامه افزایش میزان صحت طبقه‌بندی متون می‌باشد. با توجه به میزان کارهای انجام شده قبلی، افزایش کارایی طبقه‌بندهای منفرد مشکل می‌باشد، لذا رهیافت ما برای رسیدن به این هدف استفاده و بهبود روش‌های ترکیب طبقه‌بندها است.

در این پایان‌نامه برای بهبود صحت طبقه‌بندی متن، و بر مبنای روش ترکیب رأی‌گیری وزن‌دار، دو رهیافت جدید برای وزن‌دهی طبقه‌ها و طبقه‌بندها پیشنهاد شده است. رهیافت اول مبتنی بر در نظر گرفتن وزن مستقل برای هر طبقه و هر طبقه‌بند است. رهیافت دوم تعمیم رهیافت اول است. بدین شکل که برای جواب مثبت یا منفی هر طبقه‌بند در مورد هر طبقه وزن مستقلی در نظر گرفته می‌شود. برای محاسبه‌ی اوزان در هر دو رهیافت، علاوه بر الگوریتم ژنتیک، معادله تجربی خاصی هم پیشنهاد گردیده است که در زمان بسیار کمتری نسبت به الگوریتم ژنتیک اجرا می‌گردد. نتایج طبقه‌بندی بر مبنای محاسبه اوزان با استفاده از معادله‌ی پیشنهادی، کاملاً با نتایج استفاده از الگوریتم ژنتیک قابل مقایسه و حتی گاهی بهتر هم می‌باشد.

آزمایش‌ها با استفاده از طبقه‌بندهای رُکیو، نزدیک‌ترین همسایه و بیز، و سه روش انتخاب ویژگی شامل اطلاعات متقابل، خی ۲ و MCFS انجام شده است. نتایج تجربی حاصل از اعمال الگوریتم‌های ترکیب پیشنهادی بر روی مجموعه داده‌های آموزشی رایج و مقایسه با نتایج حاصل از سایر روش‌های ترکیب طبقه‌بندها، مانند رأی‌گیری وزن‌دار، عملگر میانگین وزن‌دار رتبه یافته و روش قالب تصمیم، نشان می‌دهد که رهیافت‌های پیشنهادی دقت طبقه‌بندی را بنحو چشمگیری افزایش داده‌اند. این نتایج از آزمایش بر روی چهار مجموعه داده‌های آموزشی متفاوت و رایج بدست آمده است.

**کلمات کلیدی:** طبقه‌بند، طبقه‌بند متن، ترکیب طبقه‌بندها، رأی‌گیری وزن‌دار

# فهرست مطالب

عنوان..... صفحه

## فصل ۱. مقدمه..... ۱

۱-۱	انگیزه‌ی تحقیق.....	۱
۲-۱	بیان مسأله‌ی تحقیق.....	۱
۳-۱	اهداف و رویکردها.....	۲
۴-۱	دستاوردهای پایان‌نامه.....	۲
۵-۱	ساختار پایان‌نامه.....	۳

## فصل ۲. مروری بر طبقه‌بندهای متن..... ۷

۱-۲	تفاوت طبقه‌بندی و خوشه‌یابی.....	۷
۲-۲	بخش‌های مهم طبقه‌بند.....	۸
۳-۲	طبقه‌بند نزدیک‌ترین همسایه.....	۹
۴-۲	طبقه‌بند بیز.....	۱۱
۵-۲	بازنمایی متن به صورت بردارهای عددی.....	۱۱
۶-۲	انتخاب ویژگی‌ها.....	۱۲
۱-۶-۲	انتخاب ویژگی با استفاده از روش اطلاعات متقابل.....	۱۳
۲-۶-۲	انتخاب ویژگی با استفاده از روش خی ۲.....	۱۵
۷-۲	اندازه‌گیری شباهت دو متن.....	۱۶
۸-۲	روشهای طبقه‌بندی متن.....	۱۶
۱-۸-۲	روش طبقه‌بندی نزدیک‌ترین همسایه.....	۱۷
۲-۸-۲	روش طبقه‌بندی بیز ساده.....	۲۱
۳-۸-۲	روش طبقه‌بندی رُکیو.....	۲۲

## فصل ۳. مروری بر روش‌های ترکیب طبقه‌بندها..... ۲۳

۱-۳	مقدمه.....	۲۳
۲-۳	کارایی.....	۲۴
۳-۳	نکاتی درباره‌ی ترکیب طبقه‌بندها.....	۲۴



۲۵	..... فلسفه ترکیب طبقه‌بندها	۴-۳
۲۶	..... از دید آماری	۱-۴-۳
۲۶	..... از دید محاسباتی	۲-۴-۳
۲۶	..... از دید نمایی	۳-۴-۳
۲۷	..... رهیافت‌های ساخت تیمی از طبقه‌بندها	۵-۳
۲۹	..... ترکیب و انتخاب	۶-۳
۳۰	..... انواع ترکیب‌کننده‌ی طبقه‌بندها	۷-۳
۳۰	..... انواع خروجی طبقه‌بندها	۸-۳
۳۱	..... انواع روش‌های ترکیب برچسب‌ها (خروجی سطح تجرید)	۹-۳
۳۱	..... ترکیب‌کننده‌ی رأی‌گیری	۱-۹-۳
۳۳	..... ترکیب‌کننده‌ی بیز ساده	۲-۹-۳
۳۴	..... ترکیب‌کننده‌ی با روش‌های چندجمله‌ای	۳-۹-۳
۳۹	..... ترکیب‌کننده‌ی تخمین احتمالی	۴-۹-۳
۴۲	..... مقایسه‌ی حافظه‌ی مورد نیاز روش‌های ترکیب برچسب‌ها	۱۰-۳
۴۲	..... انواع روش‌های ترکیب درجه‌ی حمایت (خروجی سطح اندازه)	۱۱-۳
۴۲	..... درجه‌ی حمایت	۱-۱۱-۳
۴۳	..... ماتریس پروفایل تصمیم	۲-۱۱-۳
۴۳	..... روش‌های ترکیب در سطح اندازه	۳-۱۱-۳
۴۳	..... روش‌های طبقه-هوشیار	۱۲-۳
۴۴	..... میانگین تعمیم داده شده [۳۲]	۱-۱۲-۳
۴۵	..... میانگین وزن‌دار رتبه‌یافته [۳۵]	۲-۱۲-۳
۴۷	..... روش‌های طبقه-بی‌تفاوت	۱۳-۳
۴۷	..... قالب تصمیم	۱-۱۳-۳

#### **فصل ۴. روش‌های پیشنهادی برای ترکیب طبقه‌بندها** ..... ۵۱

۵۱	..... رهیافت اول: در نظر گرفتن وزن‌های مستقل برای طبقات	۱-۴
	..... رهیافت دوم: در نظر گرفتن وزن‌های مستقل برای جواب‌های مثبت یا منفی	۲-۴
۵۶	..... طبقه‌بندها در هر طبقه	

## فصل ۵. نتایج و تحلیل آزمایش ..... ۶۵

۶۵	مجموعه داده	۱-۵
۶۶	مجموعه داده 20 Newsgroups	۱-۱-۵
۶۶	مجموعه داده Reuters 21578 (Top 41)	۲-۱-۵
۶۷	مجموعه داده TDT2 (Top 30)	۳-۱-۵
۶۷	مجموعه داده همشهری (ده طبقه)	۴-۱-۵
۶۸	معیارهای ارزیابی	۲-۵
۷۰	آزمایش‌ها و تحلیل نتایج	۳-۵
۷۶	آزمایش‌ها و نتایج مربوط به روش‌های پیشنهادی	۱-۳-۵

## فصل ۶. جمع‌بندی و پژوهش‌های آتی ..... ۸۳

۸۳	جمع‌بندی	۱-۶
۸۴	پژوهش‌های آتی	۲-۶

## فهرست منابع ..... ۸۷

## پیوست الف: کد الگوریتم‌های پیاده‌سازی شده ..... ۹۳

## پیوست ب: رابط کاربر نرم‌افزار توسعه یافته جهت انجام آزمایشات ..... ۱۰۳

## فهرست شکل‌ها

عنوان.....	صفحه
شکل ۱-۲: مقایسه‌ی طبقه‌بندی با خوشه‌یابی.....	۷
شکل ۲-۲: بخش‌های مختلف یک طبقه‌بند.....	۸
شکل ۳-۲: مثالی از روش نزدیک‌ترین همسایه ( $K=6$ ).....	۱۰
شکل ۴-۲: نمونه‌ای از یک اصطلاح‌نامه [۶].....	۱۹
شکل ۵-۲: نمونه‌ای از ساختار طبقه‌ها در یک اصطلاح‌نامه [۶].....	۲۰
شکل ۱-۳: ترکیب طبقه‌بندها از دید آماری [۳۱].....	۲۶
شکل ۲-۳: ترکیب طبقه‌بندها از دید محاسباتی [۳۱].....	۲۶
شکل ۳-۳: ترکیب طبقه‌بندها از دید نمایشی [۳۱].....	۲۷
شکل ۴-۳: روش‌های رأی‌گیری.....	۳۲
شکل ۵-۳: نمونه‌ای از درخت وابستگی درجه اول.....	۳۹
شکل ۶-۳: ماتریس پروفایل تصمیم.....	۴۳
شکل ۷-۳: الگوریتم روش‌های طبقه-هوشیار [۴۳].....	۴۴
شکل ۸-۳: نمایش گرافیکی الگوریتم قالب تصمیم [۳۱].....	۴۸
شکل ۱-۴: مقایسه انواع طبقه‌بندها در طبقه‌های مجموعه ۲۰ گروه خبری با معیار F1.....	۵۳
شکل ۲-۴: مقایسه طبقه‌بندی با روش‌های مختلف انتخاب ویژگی در طبقه‌های مجموعه ۲۰ گروه خبری با معیار F1.....	۵۳
شکل ۳-۴: مقایسه طبقه‌بندی با روش‌های مختلف در طبقه‌های مجموعه ۲۰ گروه خبری با معیار دقت.....	۵۴
شکل ۴-۴: مقایسه طبقه‌بندی با روش‌های مختلف در طبقه‌های مجموعه روبرتز با معیار بازیابی.....	۵۴
شکل ۵-۴: مقایسه طبقه‌بندی با روش‌های مختلف در طبقه‌های مجموعه TDT2 با معیار بازیابی.....	۵۵
شکل ۶-۴: مقایسه طبقه‌بندی با روش‌های مختلف در طبقه‌های مجموعه همشهری با معیار صحت و F1.....	۵۵
شکل ۷-۴: مقایسه جواب مثبت و منفی انواع طبقه‌بندها در طبقه‌های مجموعه‌ی ۲۰ گروه خبری.....	۵۸
شکل ۸-۴: مقایسه جواب مثبت و منفی طبقه‌بندها با روش‌های مختلف انتخاب ویژگی در طبقه‌های مجموعه‌ی ۲۰ گروه خبری.....	۵۹
شکل ۹-۴: مقایسه جواب مثبت و منفی طبقه‌بندها با روش‌های مختلف طبقه‌بندی در طبقات مجموعه‌ی روبرتز.....	۶۰

- شکل ۴-۱۰: مقایسه جواب مثبت و منفی طبقه‌بندها با روش‌های مختلف طبقه‌بندی در طبقات  
مجموعه‌ی TDT2 ..... ۶۱
- شکل ۴-۱۱: مقایسه جواب مثبت و منفی طبقه‌بندها با روش‌های مختلف طبقه‌بندی در طبقات  
مجموعه‌ی همشهری ..... ۶۲
- شکل ۵-۱: مقایسه روش‌های مختلف طبقه‌بندی و ترکیب آن‌ها بر اساس معیار ماکرو F1 ..... ۷۱
- شکل ۵-۲: مقایسه روش‌های مختلف انتخاب ویژگی در طبقه‌بند نزدیک‌ترین همسایه و ترکیب آن‌ها  
با معیار ماکرو F1 ..... ۷۲
- شکل ۵-۳: مقایسه داده‌های آموزشی متفاوت در طبقه‌بند بیز ساده و ترکیب آن‌ها با معیار ماکرو F1 ..... ۷۳
- شکل ۵-۴: مقایسه الگوریتم‌های مختلف ترکیب بر روی طبقه‌بندهای مختلف و روش‌های انتخاب  
ویژگی متفاوت ..... ۷۵
- شکل ۵-۵: مقایسه طبقه‌بندی روش‌های پیشنهادی با سایر روش‌ها با معیار ماکرو F1 ..... ۷۷
- شکل ۵-۶: مقایسه طبقه‌بندی روش‌های پیشنهادی با سایر روش‌ها با معیار میکرو F1 ..... ۷۷

# فهرست جدول‌ها

عنوان.....	صفحه
جدول ۱-۲: کلمات با بیشترین اطلاعات متقابل در مجموعه ۲۰ گروه خبری.....	۱۴
جدول ۲-۲: کلمات با اطلاعات متقابل صفر در مجموعه ۲۰ گروه خبری.....	۱۵
جدول ۱-۳: چهار رویکرد مختلف برای طراحی یک سیستم ترکیب طبقه‌بندها [۳۴].....	۲۸
جدول ۲-۳: نمونه‌ای از جدول فضای دانش-رفتار.....	۳۵
جدول ۳-۳: ماتریس انطباق طبقه‌بندهای $D_2$ و $D_3$ برای طبقه $\omega_1$ .....	۴۱
جدول ۴-۳: مقایسه‌ی روش‌های مختلف ترکیب برچسب‌ها از نظر حافظه‌ی مورد نیاز.....	۴۲
جدول ۵-۳: مقادیر مختلف پارامتر $\alpha$ برای روش میانگین تعمیم داده شده.....	۴۵
جدول ۱-۴: معیارهای مختلف طبقه‌بندی در مورد یک طبقه‌ی مشخص [۴۵].....	۵۷
جدول ۱-۵: لیست گروه‌های خبری [۲۵].....	۶۶
جدول ۲-۵: لیست طبقات مربوط به مجموعه داده‌ی همشهری.....	۶۸
جدول ۳-۵: نتایج روش‌های مختلف طبقه‌بندی و ترکیب آن‌ها بر اساس معیار ماکرو $F1$ .....	۷۲
جدول ۴-۵: نتایج روش‌های مختلف انتخاب ویژگی در طبقه‌بند نزدیک‌ترین همسایه و ترکیب آن‌ها بر اساس معیار ماکرو $F1$ .....	۷۳
جدول ۵-۵: نتایج مجموعه داده‌های آموزشی متفاوت در طبقه‌بند بیز ساده و ترکیب آن‌ها بر اساس معیار ماکرو $F1$ .....	۷۴
جدول ۶-۵: نتایج الگوریتم‌های مختلف ترکیب بر روی طبقه‌بندهای مختلف و روش‌های انتخاب ویژگی متفاوت.....	۷۵
جدول ۷-۵: نتایج طبقه‌بندی روش‌های پیشنهادی و سایر روش‌ها با معیار ماکرو $F1$ .....	۷۸
جدول ۸-۵: نتایج طبقه‌بندی روش‌های پیشنهادی و سایر روش‌ها با معیار میکرو $F1$ .....	۷۸
جدول ۹-۵: نتایج طبقه‌بندی روش‌های پیشنهادی و سایر روش‌ها با معیار ماکرو دقت.....	۷۹
جدول ۱۰-۵: نتایج طبقه‌بندی روش‌های پیشنهادی و سایر روش‌ها با معیار ماکرو بازیابی.....	۸۰
جدول ۱۱-۵: نتایج طبقه‌بندی روش‌های پیشنهادی و سایر روش‌ها با معیار صحت.....	۸۰

## فصل ۱. مقدمه

### ۱-۱ انگیزه‌ی تحقیق

با توجه به گسترش اینترنت و افزایش حجم مطالب متنی، نیاز به روش‌های مدیریت متون و مطالب بطور چشمگیری افزایش یافته است [۱]. طبقه‌بند متن یکی از ابزارهایی مطرح در هوش مصنوعی است که می‌تواند در این زمینه مؤثر باشد. طبقه‌بندی متن کاربردهای فراوانی دارد و از جمله کاربردهای آن می‌توان به موتورهای جستجو [۲-۴]، سیستمهای فیلترینگ [۳]، سیستمهای دسته‌بندی متون [۵-۸]، کتابخانه‌های دیجیتال [۶]، دسته‌بندی خبرهای دریافت شده در گروه‌های خبری [۴، ۹]، تشخیص هزینه‌ها در سیستمهای پست الکترونیک [۱۰، ۱۱] و بسیاری کاربردهای دیگر اشاره کرد.

### ۲-۱ بیان مسأله‌ی تحقیق

با توجه به افزایش چشمگیر حجم مطالب متنی در سال‌های اخیر، و با توجه به کاربردهای مهم طبقه‌بندی متن، نیاز به طبقه‌بندهای متنی که دارای صحت طبقه‌بندی بالایی باشند اهمیت زیادی پیدا کرده است. خطا در طبقه‌بندی متن در بعضی از کاربردها هزینه‌های زیادی را ممکن است داشته باشد. برای نمونه اگر در یک سیستم پست الکترونیک به اشتباه یک نامه‌ی مهم به عنوان هزینه‌نامه تشخیص

داده شود و کاربرد متوجه نامه‌ی مذکور نشود ممکن است هزینه‌ی گزافی برای کاربر یا شخص ارسال‌کننده داشته باشد. با توجه به پیچیدگی حوزه‌ی پردازش متن، هنوز بهبود میزان صحت طبقه‌بندی متن از مسائل مطرح در این حوزه می‌باشد.

### ۱-۳ اهداف و رویکردها

با توجه به مطالب ذکر شده درباره‌ی اهمیت صحت طبقه‌بندی متن، مهمترین هدف در این پایان‌نامه رسیدن به میزان صحت بیشتر در طبقه‌بندی متن می‌باشد. برای این منظور از روش‌های ترکیب در طبقه‌بندی متن استفاده شده است. روش‌های رأی‌گیری، عملگر میانگین وزن‌دار رتبه یافته و روش قالب تصمیم از جمله روش‌های استفاده شده برای رسیدن به هدف اصلی در این پایان‌نامه است. همچنین تاثیر نحوه انتخاب ویژگی بر ترکیب طبقه‌بندها بررسی شده است.

### ۱-۴ دستاوردهای پایان‌نامه

در راستای رسیدن به اهداف پایان‌نامه، دستاوردهای مهمی حاصل شده است که در ادامه بصورت مختصر هر یک از این دستاوردها شرح داده می‌شوند و در فصل‌های بعد به تشریح جزئیات آنها پرداخته خواهد شد.

همانطور که در بخش‌های قبل نیز اشاره شد، اصلی‌ترین هدف این پایان‌نامه رسیدن به میزان صحت بیشتر در طبقه‌بندی متن است. در این پایان‌نامه، ما به بررسی روش‌های مختلف طبقه‌بندی متن پرداخته‌ایم. همچنین روش‌های مختلف ترکیب طبقه‌بندها نیز بررسی شده است. مبنای روش‌های پیشنهادی ما در این پایان‌نامه، روش رأی‌گیری وزن‌دار است. برای وزن‌دهی دو رهیافت جدید ارائه داده‌ایم. رهیافت اول مبتنی بر در نظر گرفتن وزن مستقل برای هر طبقه از هر طبقه‌بند است. رهیافت دوم تعمیم رهیافت اول است. بدین شکل که برای جواب مثبت یا منفی هر طبقه‌بند و برای هر طبقه

وزن مستقلى در نظر گرفته شود. پس از پياده‌سازى چندين طبقه‌بند، با در نظر گرفتن اوزان تطبيقى وابسته به طبقه، طبقه‌بند و همچنين با توجه به خروجى آن در روش‌هاى تركيب طبقه‌بندها، به نتايج بهترى در صحت طبقه‌بندى متن رسيديم.

دستاورد ديگرى كه تركيب طبقه‌بندها به همراه دارد، افزايش پايدارى طبقه‌بندى مى‌باشد. زيرا چندين طبقه‌بند به صورت موازى با هم عمل طبقه‌بندى را انجام مى‌دهند و در صورتى كه بعضى از آنها دچار مشكل شوند، ساير طبقه‌بندها مى‌توانند يك جواب مناسب را توليد كنند و سيستم متوقف نخواهد شد. از طرف ديگر چون طبقه‌بندها بصورت موازى مى‌توانند اجرا شود، با استفاده از چندين پردازشگر و با استفاده از تكنيك‌ها پردازش موازى مى‌توان در زمان طبقه‌بندى كل سيستم صرفه‌جويى قابل ملاحظه‌اى كرد [۱۲].

## ۱-۵ ساختار پايان‌نامه

اين پايان‌نامه در شش فصل تنظيم گرديده است كه رئوس مطالب و موضوعات اصلى هر فصل در ذيل بطور خلاصه بيان گرديده است:

- **فصل اول** به انگيزه، اهميت و شرح موضوع تحقيق مى‌پردازد. همچنين ساختار كل پايان‌نامه و خلاصه مطالب ارائه شده در اين تحقيق نيز بيان مى‌گردد.
- **فصل دوم** مطالب پيش‌زمينه‌اى را كه براى درك روش‌هاى ارائه شده در پايان‌نامه ضرورى است، بيان مى‌كند. از آنجايكه هدف اين پايان‌نامه بهبود طبقه‌بندى متن است، لازم است مقدمه‌اى درباره‌ى طبقه‌بندى و روش‌هاى مختلف آن ارائه شود. چون در طبقه‌بندى متن از



روش‌های نزدیک‌ترین همسایه، بیز ساده و رُکیو<sup>۱</sup> بیشتر استفاده می‌شود، و در آزمایشات این پایان‌نامه از این سه روش استفاده شده است، مروری بر آنها انجام داده‌ایم.

برای استخراج ویژگی‌ها از متن و انتخاب ویژگی‌های مهم، الگوریتم‌های مختلفی در مقالات پیشنهاد شده است. ما از دو روش معروف اطلاعات متقابل<sup>۲</sup> [۱۳] و  $\chi^2$  [۱] و همچنین از یک روش جدید به نام MCFS [۱۴] استفاده کرده‌ایم. در ادامه‌ی این فصل روش‌های رُکیو و بیز ساده [۱۵] و نزدیک‌ترین همسایه [۱] برای طبقه‌بندی متن بررسی شده است.

## • فصل سوم از آنجاییکه رهیافت ما در این پایان‌نامه استفاده از روش‌های ترکیب طبقه‌بندها

است، این فصل را به مروری بر نحوه‌ی ترکیب طبقه‌بندها و مطالب پیش‌زمینه‌ای که برای درک روش‌های ترکیب طبقه‌بندها ضروری است، اختصاص داده‌ایم [۱۶, ۱۷]. روش‌هایی که مرور خواهند شد عبارتند از رأی اکثریت، رأی اکثریت وزن‌دار، ترکیب بیز ساده، روش‌های طبقه-هوشیار<sup>۴</sup> و روش‌های طبقه-بی‌تفاوت<sup>۵</sup>.

## • فصل چهارم به بررسی راهکارهای پیشنهادی برای بهبود صحت طبقه‌بندی متن می‌پردازد.

رهیافت ما استفاده از روش‌های ترکیب طبقه‌بندها است. ما در این پایان‌نامه دو رهیافت جدید در روش ترکیب با استفاده از رأی‌گیری وزن‌دار ارائه داده‌ایم. رهیافت اول مبتنی بر در نظر گرفتن وزن مستقل برای هر طبقه از هر طبقه‌بند است. رهیافت دوم تعمیم رهیافت اول است. بدین شکل که برای جواب مثبت یا منفی هر طبقه‌بند و برای هر طبقه وزن مستقلی در نظر

---

<sup>1</sup> Rocchio

<sup>2</sup> Mutual Information

<sup>3</sup> Chi-2

<sup>4</sup> Class-Conscious

<sup>5</sup> Class-Indifferent

گرفته شود. برای محاسبه وزن در هر دو رهیافت، معادله‌ی تجربی نیز پیشنهاد شده است. البته از سایر روش‌های هوشمند مانند الگوریتم ژنتیک نیز برای محاسبه وزن‌ها بهره برده شده است.

#### • **فصل پنجم** نتایج حاصل از پیاده‌سازی روش‌های طبقه‌بندی متن که در فصل سوم و چهارم

بیان شد را ارائه کرده‌ایم. با تقسیم داده‌های آموزشی به سه دسته و استفاده از سه روش مختلف انتخاب ویژگی (اطلاعات متقابل،  $\chi^2$  و MCFS)، و سه روش طبقه‌بندی متن (رُکیو، بیز ساده و نزدیک‌ترین همسایه)، مجموعه متنوعی از طبقه‌بندهای پایه را جهت انجام عملیات ترکیب آماده نمودیم. سپس با استفاده از روش‌های رأی‌گیری و میانگین وزن‌دار رتبه‌یافته و قالب تصمیم، نتایج حاصل از طبقه‌بندهای متن را با هم ترکیب کردیم و به نتایج بهتری رسیدیم. در ادامه روش‌های پیشنهادی که دو رهیافت جدید برای محاسبه‌ی وزن‌ها در روش رأی‌گیری بود را بکار بردیم. برای محاسبه وزن در هر دو رهیافت، هم از معادله‌ی تجربی پیشنهاد شده در فصل چهارم استفاده شده و هم از الگوریتم ژنتیک بهره برده شده است. نتایج تجربی نشان می‌دهد که رهیافت پیشنهاد شده، دقت طبقه‌بندی را نسبت به سایر روش‌های رأی‌گیری و همچنین سایر روش‌های ترکیب در مجموع افزایش داده است. این نتایج از آزمایش بر روی چهار داده آموزشی متفاوت بدست آمده است.

#### • **فصل ششم** در این فصل راهکارهای ارائه شده در این پایان‌نامه بصورت خلاصه بررسی شده

و در ادامه پیشنهادهای برای توسعه و بهینه‌سازی هرچه بیشتر این سیستم ارائه می‌شود.

## فصل ۲. مروری بر طبقه‌بندی‌های متن

طبقه‌بندی‌ها از مباحث مطرح در بازساخت الگو می‌باشند. روش‌های بسیار زیادی برای طبقه‌بندی ارائه شده است. هدف اصلی از ارائه‌ی روش‌های مختلف، افزایش صحت طبقه‌بندی می‌باشد. یک ایده‌ی مهم که در این زمینه مطرح می‌باشد ترکیب چند طبقه‌بند برای رسیدن به نتایج بهتر می‌باشد. اما در ابتدا باید با نحوه کار طبقه‌بندی‌ها آشنا شد، لذا در این فصل به مرور اجمالی طبقه‌بندی‌های پرکاربرد می‌پردازیم.

### ۱-۲ تفاوت طبقه‌بندی<sup>۱</sup> و خوشه‌یابی<sup>۲</sup>



شکل ۱-۲: مقایسه‌ی طبقه‌بندی با خوشه‌یابی

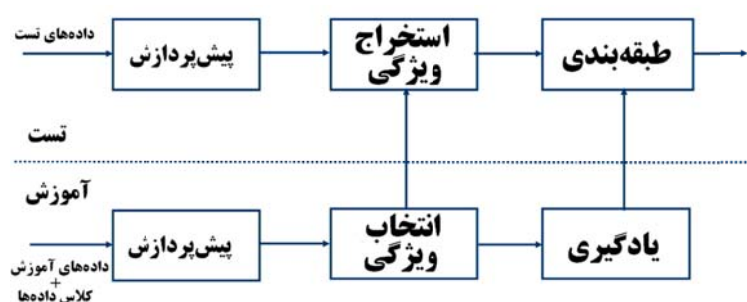
<sup>1</sup> Classification

<sup>2</sup> Clustering

در طبقه‌بندی، طبقات از قبل شناخته‌شده هستند. هدف یک طبقه‌بند این است که تشخیص دهد ورودی به کدام طبقه تعلق دارد. طبقه‌بندها به صورت با سرپرست<sup>۱</sup> آموزش داده می‌شوند. اما در خوشه‌یابی، طبقه‌ها لزوماً از قبل شناخته‌شده نیستند و الگوریتم خوشه‌یابی با توجه به داده‌های ورودی، تعداد طبقه‌ها و نحوه‌ی طبقه‌بندی را تشخیص می‌دهد. در واقع می‌توان این الگوریتم‌ها را بدون سرپرست<sup>۲</sup> نامید. تفاوت طبقه‌بندی و خوشه‌یابی در شکل ۱-۲ نمایش داده شده است.

## ۲-۲ بخش‌های مهم طبقه‌بند

در شکل ۲-۲ بخش‌های مهم یک طبقه‌بند را مشاهده می‌کنید.



شکل ۲-۲: بخش‌های مختلف یک طبقه‌بند

پیش‌پردازش<sup>۳</sup>: این بخش داده‌های ورودی را به شکلی که برای طبقه‌بندی مناسب باشد بازنمایی می‌کند [۴].

استخراج ویژگی<sup>۴</sup>: در این مرحله، بخشی از اطلاعات که برای طبقه‌بندی مهم نیست، حذف شده و اطلاعات مفید برای طبقه‌بندی تحت عنوان ویژگی استخراج می‌شود. این کار باعث می‌شود تا ضمن حفظ کارایی طبقه‌بند، پیچیدگی محاسباتی کاهش یابد [۴, ۱۸].

<sup>1</sup> Supervised  
<sup>2</sup> Unsupervised  
<sup>3</sup> Preprocessing  
<sup>4</sup> Feature Extraction