# Asistencia para la categorización de documentos

Álvaro Moncada – alvaro.moncada @javeriana.edu.co
Álvaro Pérez – perez.alvaro @javeriana.edu.co
Edwin Turizo – edwin.turizo@javeriana.edu.co

**Resumen.** El presente documento tiene como intención abordar la problemática de la categorización de documentos mediante el uso de lógica difusa (LD). Puntualmente, el objetivo consiste en utilizar un algoritmo basado en clustering difuso que asista en la segmentación de documentos enmarcado en un sistema inteligente basado en agentes racionales. Dado el alcance de este documento, se propone una solución para la categorización de noticias utilizando el algoritmo Fuzzy C-Means (FCM) y adicionalmente se tendrán procesos de tratamiento de los datos de entrada como *tokenización* y *stop-words*. El set de datos utilizado consta de 2225 noticias de la BBC clasificadas en 5 categorías para finalmente determinar el desempeño del modelo frente a la problemática abordada.

**Palabras Clave:** Clasificación de documentos, Lógica Difusa, Fuzzy C-Means, Clustering.

## I. Contextualización del problema

El descubrimiento de evidencia digital o E-Discovery, permite a investigadores o abogados identificar información almacenada electrónicamente relacionada con el objetivo de un caso. Luego de identificar aquellos archivos relevantes mediante el uso de palabras clave inicia la etapa de revisión que, usualmente, consume muchos recursos y toma mucho tiempo debido al gran volumen de información a revisar. Por lo que es relevante la implementación de técnicas que reduzcan estos tiempos y permitan a los revisores enfocar sus esfuerzos para identificar la mayor cantidad de información relevante ágilmente.


- Incremento de costos debido al alto número de revisores involucrados
- Posibles diferencias en los criterios de clasificación

- Disminución de cantidad de personas involucradas y menor costo asociado
- Homogeneidad en los criterios de clasificación

**Ilustración 1. Comparación de modalidades para la revisión de documentos**

*. https://www.kaggle.com/c/learn-ai-bbc/data

# II. Análisis del estado del arte

Dado nuestro alcance, nos enfocaremos en indagar el estado del arte para soluciones relacionadas con la clasificación de noticias o de cualquier texto en general. Este tipo de soluciones constan de dos (2) etapas principales, una etapa inicial de alistamiento de los datos que consta de un **preprocesamiento y limpieza del dataset** y un posterior entrenamiento de los **algoritmos utilizados** para la clasificación de documentos [1], dado el alcance de este documento, se limitará a lógica difusa y más específicamente a Fuzzy c means.

Con respecto a Fuzzy C-Means (FCM), se decidió por esta técnica puesto que Thaung Thaung Win y LinMon [3] presentaron como utilizar dicha técnica para clasificar documentos de texto con clústeres o grupos predefinidos como lo es en nuestro caso, en el cual el dataset ya cuentan con las categorías definidas de las noticias. Adicionalmente, en el caso de estudio realizado por los autores previamente mencionados afirman que es una técnica eficiente y con la capacidad necesaria para manejar grandes volúmenes de datos. FCM consiste en agrupar las noticias en clusters apoyados con la determinación de si una noticia pertenece a un grupo o a otro a través del grado de pertenencia entre 0 y 1. Así mismo, cada noticia tiene características relevantes y diferenciadores que en este caso son las palabras en donde Xi es una noticia o punto en m-dimensiones y m cada palabra que compone la noticia.

FCM utiliza como fórmula de optimización

$$J_m(U,v)=\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m\left\|x_k-v_i\right\|_A^2$$ , donde $U=[u_{ik}]\in M_{fc}$

Es una matriz de partición difusa de X, uik es la membresía del k-esimo punto de datos de la I-esima clase

$V=[v_1,v_2,v_3,...,v_c], v_i\in \mathbb{R}^n$ es el vector de prototipos de clusters que son predeterminados.

$D_{ikA}^2=\left\|x_k-v_i\right\|_A^2=(x_k-v_i)^T A(x_k-v_i)$ es la distancia Euclidiana entre cada clúster.

Los parámetros en FCM son:

- Número de clústeres: la cantidad de clústeres es representado por c, en este caso es predeterminado por las 5 categorías.
- Parámetro difuso: Representado por m, es el encargado definir qué tan difuso es la partición entre los grupos. Entre menor este valor se dice que es más difuso.
- Número de iteraciones: cantidad de veces que se repetirá el proceso para encontrar la mejor solución.

La función de evaluación de precisión del modelo esta dado por la medida F-Measure, que está dada por la siguiente ecuación.

$$F - Measure_i = \frac{2 \bullet \Pr ecision_i \bullet \operatorname{Re} call_i}{\Pr ecision_i + \operatorname{Re} call_i}$$

Este valor se calcula para cada cluster i, con el fin de determinar el error de clasificación que se puede tener en los grupos generados.

En donde la Precisión se calcula con la fórmula $\mathbf{Precision_i = CN_i /JN_i}$,

Y el Recall se calcula con la fórmula $\mathbf{Recall_i = CN_i /RN_i}$, definiéndose:
CNi: Número de documentos juzgados de forma correcta en el cluster i,
JNi: Número de documentos juzgados en el cluster i,
RNi: Número de documentos pre-etiquetados en el cluster i
En cuanto al F-Measure General, este se define de la siguiente manera:

$$Overall \;\; F - Measure = \frac{\sum_{i=1}^{C}(RN_i \bullet F - Meausre_i)}{\sum_{i=1}^{C} RN_i}$$

La inicialización del proceso se realiza de la siguiente manera Cano, C. Cano, E. [3] se debe seleccionar el número de clusters. Posteriormente se determinan los centroides.

$$V_i^{(l)} = \frac{\sum_{k=1}^{n} (u_{ik}^{(l-1)})^{m`} x_k}{\sum_{k=1}^{n} (u_{ik}^{(l-1)})^{m`}}$$

Paso 2: se calcula las distancias

$$D_{ik}^2 = \|z_k - v_i\|^2 = \sqrt{(z_k - v_i)^T A (z_k - v_i)}, 1 \le i \le c, 1 \le k \le N$$

Paso 3: reorganización nuevamente de la matriz de partición difusas con las siguientes reglas

$$\text{Si } D_{ik} > 0 \; para \; 1 \le i \le c, 1 \le k \le N,$$

$$\mu_{ik}^{(1)} = \frac{1}{\sum \left( \frac{D_{IK}}{D_{JK}} \right)^{2/m-1}}$$

entonces:

$$\text{De otra forma: } \mu_{ik}^{(1)} = 0$$

$$\text{Hasta : } \left| U^{(l)} - U^{(l-1)} \right| < \varepsilon$$

En un caso de estudio, Wei Wang (2008) [4] propuso un marco para combinar el algoritmo de agrupación en clústeres difuso (FCM) y los métodos de selección de características supervisadas basados en el algoritmo Maximización de Expectativas (EM). La estadística χ2 (CHI CUADRADO) mide la dependencia entre los términos textuales de los documentos y el cluster con el fin de agrupar los documentos similares e ir incrementando el número de documentos que son clasificados allí. La medida estadística χ2 solo le importa si el documento tiene el término, pero no le importa la frecuencia del término en el documento, esto debido a que se toma como factor crucial la pertenencia de un término que pueda representar a un cluster.

En este caso se utilizó un dataset "4-News" de noticias que contenía 3997 registros y 4 categorías pre-definidas. Para la validación se eligió F-Measure para evaluar el algoritmo de clustering. Esta medida se utiliza habitualmente para validar la eficacia de la agrupación en clústeres generados. Se obtuvo un valor de F-Measure general de 70%. Lo que se aplicó de este paper con respecto a nuestra problemática, es la importancia de tener indicadores relacionados con la precisión de la generación de los clusters como lo es la medida del F-Measure, adicionalmente, la generación de la matriz binaria de la existencia de un término con respecto a la noticia de forma que pueda representar con 1s y 0s sí una palabra tiene mayor relevancia en la agrupación de la noticia en un cluster categorizado.

Por otro lado, para la fase de alistamiento de la información, vemos diferentes aproximaciones a esta etapa. La extracción de descriptores (*features*) para la clasificación de texto es especialmente importante para la clasificación de texto, ya que tiene esta problemática posee dos características relevantes, su alta dimensionalidad y la presencia de palabras "ruidosas", es decir que son frecuentes, pero no aportan a la tarea de clasificación. En general un texto puede ser representado de dos maneras. La primera es llamada *bag-of-words,* en la que un documento es representado como un set de palabras junto con su frecuencia asociada en el documento. Este tipo de representación es independiente del orden de las palabras. La segunda manera consiste en representar el texto como *strings,* en este caso se tiene en cuenta la secuencia de las palabras.[1]

Una de las maneras más comunes de extracción de descriptores tanto para modelos supervisados como no supervisados es el descarte de *stop-words* y *stemming.* El primero consiste en la exclusión de palabras que no son específicas para las diferentes categorías analizadas. Por otro lado, en *stemming* las diferentes formas de una palabra son consolidadas en una sola. Para nuestro caso, vamos a utilizar *stop-words*, junto con *tokenization* el cual consiste en reducir un texto a unidades individuales, que para nuestra aplicación serán palabras. [1]

# III. Metodología

**CRISP DM**

A continuación, se presenta una descripción de cada fase y lo realizado dentro de cada subproceso que abarca la metodología CRIPS-DM 1.0 [2]:

1. **Entendimiento del negocio**
   En esta fase se busca establecer las metas y objetivos a cumplir con el proyecto de inteligencia artificial basado en los métodos de Algoritmos Genéticos. En este caso lo que se busca es clasificar noticias en un conjunto de categorías ya definidas con el fin de automatizar el proceso manual a las personas interesadas.
2. **Entendimiento de los datos**
   En esta fase se busca analizar con profundidad el *dataset* a utilizar, identificar la pertinencia de este, y los atributos y características propios, esto es, hacer un barrido sí se cumplen ciertos criterios definidos para lo que se

necesita dentro del problema y con esto, la calidad con que vienen para su posterior utilización.

Nuestra aplicación al estar relacionada con el tratamiento de texto tiene como características que es data no estructurada y tiene una alta dimensionalidad y es disperso, es decir que no todas las palabras van a aparecer en todos los documentos. Así mismo, La calidad de la data es aceptable dado que no se identificaron campos faltantes o caracteres no legibles.

3. **Preparación de los datos**

   Se utilizó un dataset de 2.225 noticias de la BBC distribuida en 5 categorías que son: Deporte, Entretenimiento, Negocio, Tecnología y Política. Dicha base cuenta tres columnas, una es el texto de la noticia, el número de noticia y la categoría a la que pertenece, importante resaltar que se evidencia una distribución similar de las cinco categorías a lo largo del dataset. Para el procesamiento de los datos se realizó una tokenización, la cual consiste en dividir las palabras del texto de la noticia como un solo atributo del registro o segmentar las frases contenidas en la noticia, posteriormente se definió la longitud de las palabras a tener en cuenta en el modelado, en este caso se utilizaron aquellas palabras con 3 o más letras. Luego, se aplicó stopwords, lo cual permite identificar aquellas palabras que no son relevantes para clasificar el texto, suelen ser palabras frecuentes que no ayudan a segmentar la noticia a una categoría. Por último, se usó stemming con el fin de reducir palabras que tienen una misma raíz, es decir, que debido a la manera que se conjugan pueden agruparse en la misma categoría con el fin de identificar patrones de una manera más rápida y sencilla.

4. **Modelado**

   Consiste en la aplicación de la técnica a utilizar, en este caso algoritmos genéticos y para evaluar su efectividad utilizaremos como medida de desempeño el fitness y más específicamente el F-Measure, el cual consiste en combinar las medidas de precision y recall.

   Se utiliza dicha medida debido a la naturaleza del problema a solucionar se debe evaluar por igual que la calidad y capacidad de clasificar correctamente la noticia. Adicionalmente se dividirá la cantidad de features o palabras obtenidas en el preprocesamiento que son más relevantes o ayudan a clasificar las noticias y agregarlas al Rapidminer, la cual es la herramienta a utilizar en esta ocasión puesto que dicha herramienta cuenta con la opción de Fuzzy C means.

5. **Evaluación**

   En primera instancia se definió como criterio de evaluación del modelo la clasificación de las noticias en alguna de las cinco categorías correctamente y como medida se seleccionó el F-Measure. Para encontrar el mejor F-Measure se manipularon los diferentes parámetros de Rapidminer como son y se utilizaran diferentes parámetros como son el difuso, information gain, cantidad de palabras. Iteraciones y tipo de distancia. se evaluará si aplicar o no stopwords tiene algún impacto en el modelo.

6. **Despliegue**
   En esta fase se busca llevar el modelo a producción de forma que utilice el modelo con datos reales, sin embargo, al ser un proyecto de investigación en el que lo que se busca es llevar a cabo la aplicación y experimentación de cada una de las técnicas principales en los distintos módulos, los resultados son presentados por medio de los distintos papers publicados.

# IV. Modelo para la aplicación de los conceptos y algoritmo de la técnica IA en el caso desarrollado

Para la implementación de esta solución se utilizó el módulo Information Selection de RapidMiner, el cual nos permitió implementar Fuzzy c-means para realizar la clasificación del dataset de las 2225 noticias.

Para esto, se implementó una etapa de preprocesamiento de las noticias para pulir la información, para luego alimentar el algoritmo de Fuzzy c-means. A continuación, se detallan las secciones antes expuestas.

### Preprocesamiento

Como se ha mencionado anteriormente, se utilizó un *dataset* de 2225 noticias de la BBC que se encuentran clasificadas en 5 categorías. Antes, de poder analizar los datos mediante algoritmos de lógica difusa, es importante realizar un preprocesamiento del texto de las noticias, el cual permita una mejor interpretación de los datos. Para esto, se implementó el siguiente modelo.

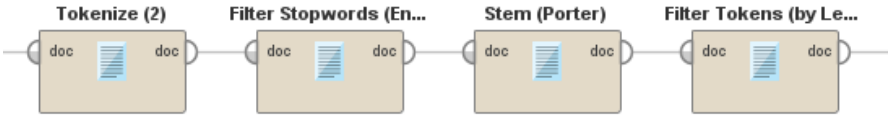



**Ilustración 2. Preprocesamiento de noticias**

Inicialmente, se **tokenizó** el texto de la notica generando un diccionario completo de todas las palabras en el dataset. Luego, se descartan los ***stop-words*** utilizando el vocabulario en inglés, posteriormente se aplica la técnica de ***stemming*** con el algoritmo Porter para obtener únicamente las raíces de las palabras y eliminar variantes gramaticales, para finalmente filtrar aquellas palabras que tengan menos de 2 letras, ya que estas generalmente no aportan para la clasificación.

Esto da como resultado un vector de palabras la cual se genera utilizando *"Binary Term Ocurrences"*. Este vector de palabras cuenta con valores de 1s y 0s en las columnas de las palabras en cada noticia (filas), donde 1 significa que esa palabra está presente y 0 que no.

Finalmente, seleccionamos las 100 palabras con el mayor Information Gain, este paso fue necesario incluirlo debido a que si utilizábamos todo el universo de palabras la precisión de la clasificación se disminuía.

### Fuzzy C-Means

A partir de la revisión del estado del arte, optamos por utilizar el algoritmo Fuzzy C-Means de la herramienta RapidMiner, la cual permite dar solución a la problemática de clasificación de documentos brindando un grado de pertenencia difuso a los grupos de cada documento y mediante el cálculo del grupo con mayor grado de pertenencia defuzzyfica. Para esto se utilizó el siguiente diagrama de flujo.

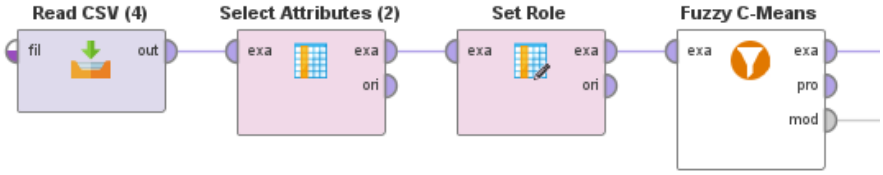



**Ilustración 3. Modelo clasificador utilizado**

Este flujo inicia con la importación de la matriz de [11131 x 2225] palabras vectorizadas. Luego, se seleccionan los 100 atributos o palabras con mayor Ingormation Gain, para luego pasar al modelo Fuzzy C-Means, pero antes fue necesario asignar un rol de clasificación a la columna que contiene los labels, esto para mayor facilidad de representación de los resultados.

Luego de revisar el estado del arte se inició una serie de iteraciones para identificar los parámetros que optimizaran la salida del clasificador. Para esto se utilizaron 5 clusters, dado que el dataset de noticias está dividido en 5 categorías; el número de iteraciones se mantuvo por defecto en 50; el parámetro de fuzzyness, indica qué tan difusos van a ser los conjuntos, se manejó el valor por defecto de 2, es decir si este aumenta la mayoría de los documentos quedaban clasificados en un solo grupo; otro parámetro fue MinGain, el cual se varió hasta encontrar aquel que diera mejores resultados, que para nuestro caso fue de 5e-4.

A continuación, una imagen de la configuración del que sería nuestro modelo base.

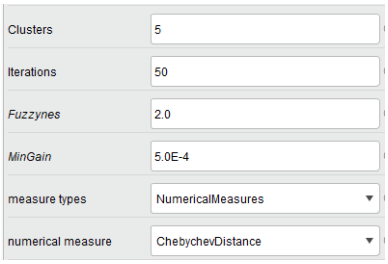



**Ilustración 4. Configuración de Fuzzy C-Means**

Para el cálculo del accuracy de nuestro clasificador, el estado del arte utilizó el F-Measure, el cual se calcula en base al Precición y el Recall definidos de la siguiente manera:

$$\mathbf{Precision}_i = \mathbf{CN}_i / \mathbf{JN}_i \,,\; \mathbf{Recall}_i = \mathbf{CN}_i / \mathbf{RN}_i \,,$$

$$F - Measure_i = \frac{2 \bullet \mathbf{Pr}\, ecision_i \bullet \mathbf{Re}\, call_i}{\mathbf{Pr}\, ecision_i + \mathbf{Re}\, call_i} \,,$$

Donde CN corresponde al número de documentos correctamente clasificados para la clase i; JN es el número de documentos clasificados en la clase i; y RN el número de documentos reales de la clase o categoría.
Por otro lado, luego de realizar pruebas con varios tipos de medidas (Measure Types) la que mejores resultados dio para nuestra problemática fue Chebychev Distance.

# V. Protocolo experimental y análisis de los resultados obtenidos

Como se introdujo en la sección anterior, nuestro modelo base esta parametrizado de la siguiente manera:

| ***Parámetro*** | **Valor** |
|---|---|
| Número Clusters | 5 |
| Fuzzyness | 2.0 |
| Min Gain | 0.0005 |
| Iteraciones | 50 |
| Métrica Distancia | Chebyshev |
| F-Measure General | 38.0% |

El proceso se trabajó con las 2225 noticias y se acortaron a únicamente 100 palabras teniendo como base su information gain obtenido de la matriz binaria de términos, que establecía sí el termino textual aparecía en cada noticia y en base a esto se escogieron los mejores 100 términos que daban mayor significado en la categorización.

A continuación, se presentan los resultados obtenidos a partir de la variación de las variables independientes, factores controlables y el respectivo valor de F-Measure para cada categoría en el proceso de Fuzzy C-Means.

Variación *valor de Fuzzyness de FCM*:

| **Fuzzyness** | **F-Measure Politics** | **F-Measure Tech** | **F-Measure Sport** | **F-Measure Business** | **F-Measure Entertainment** | **F-Measure General** |
|---|---|---|---|---|---|---|
| 2 | 58.9% | 44.6% | 48.6% | 2.8% | 31.0% | 38.0% |
| 2.5 | 43.9% | 26.2% | 0.0% | 34.2% | 55.1% | 32.18% |
| 3 | 24.6% | 17.6% | 0.0% | 26.5% | 24.1% | 18.95% |
| 5 | 47.9% | 29.4% | 0.4% | 44.5% | 55.6% | 36.08% |
| 10 | 40.4% | 35.2% | 2.2% | 29.8% | 57.5% | 32.99% |

A partir de estos resultados, podemos ver que los valores grandes de Fuzzyness difuminarán las clases y esto se debe a que, con un valor grande de m, todos los elementos tienden a pertenecer a todos los grupos por lo cual al ser un cluster por categoría [5], se ubican documentos de noticias en una categoría que no es su correspondiente y esto incrementa los errores de validación y contribuye a la disminución de F-Measure.

Variación *Minimun Gain* de FCM:

| **Min Gain** | **F-Measure Politics** | **F-Measure Tech** | **F-Measure Sport** | **F-Measure Business** | **F-Measure Entertainment** |
|---|---|---|---|---|---|
| 1 | 54.8% | 25.3% | 5.1% | 40.8% | 47.1% |
| 0.5 | 54.8% | 25.3% | 5.1% | 40.8% | 47.1% |
| 0.0005 | 58.9% | 44.6% | 48.6% | 2.8% | 31.0% |
| 0.00005 | 58.9% | 44.6% | 48.6% | 2.8% | 31.0% |
| 0.000005 | 58.9% | 44.6% | 48.6% | 2.8% | 31.0% |

La variación del Min Gain tiene un leve impacto en los valores del F-Measure, siempre y cuando se altere grandemente el orden de magnitud al valor numérico de la mínima ganancia, puesto que, al tener un menor valor de este parámetro, no se tiene una aceptación al cambio de las iteraciones que se establecen, esto es, los valores de las funciones de membresía no se ven afectados puesto que no se tiene una ganancia considerable. También se observa que, a un mayor número de Min Gain, el valor del F-Measure tiende a aumentar hasta llegar al valor de 1 que es el máximo valor que soporta este parámetro.

Adicionalmente, se realizaron variaciones al número de iteraciones, sin embargo, ninguna de los clústeres presentó variación, y tampoco la asignación de cada noticia a los grupos se vio afectada, es decir el F-Measure del modelo base para cada categoría no se alteró.

Aplicación de *stopwords* en el preprocesamiento de los datos

| **Aplicación Stop-Words** | **F-Measure Politics** | **F-Measure Tech** | **F-Measure Sport** | **F-Measure Business** | **F-Measure Entertainment** |
|---|---|---|---|---|---|
| Con Stop-Words | 58.9% | 44.6% | 48.6% | 2.8% | 31.0% |
| Sin Stop-Words | 57.1% | 5.9% | 57.6% | 49.0% | 1.7% |

Por otro lado, el efecto de no aplicar *stop words* durante el preprocesamiento si afecta mucho en algunos casos en el F-Measure del clustering, por ejemplo, para las noticias de Tech y Business se observa una disminución aproximada de 40 puntos porcentuales. Este comportamiento se puede deber a que al aplicar Stop words ciertas palabras que existen en muchas noticias y se repiten un número grande de veces, hace que se tenga un mayor information gain a esos términos que en la vida real no

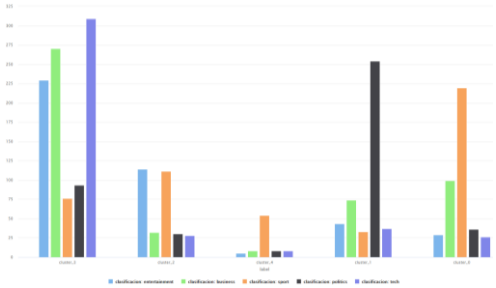
tienen significado para la segmentación de la noticia en una categoría que si tiene cercanía con términos específicos del tópico.

Matriz de membresía difusa

Con respecto a los resultados, se obtuvo la matriz de funciones de membresía que muestra los valores de pertenencia de cada noticia al Clustering generado representado por una categoría de la noticia. A continuación, se muestra un resumen de las primeras 2 noticias con sus respectivos valores de pertenencia que están en un rango de [0, 1] y en cada noticia, la columna con el mayor valor es la categoría que se le es asignada a esa noticia.

| News | Confidence (Business) | Confidence (Politics) | Confidence (Entertainment) | Confidence (Sport) | Confidence (Tech) |
|---|---|---|---|---|---|
| 1 | 0.20075001960032254 | 0.19915772634984488 | 0.19949497697926327 | 0.20095198286408616 | 0.1996452942064832 |
| 2 | 0.20182854772947437 | 0.19920789489995047 | 0.19904171983525695 | 0.1995795035294857 | 0.2003423340058325 |

A continuación, se muestra la gráfica que representa la cantidad de noticias totales que se agruparon en cada cluster, se observa que en un mismo clusters se tienen noticias que no corresponden a la categoría por lo que se intuye que la segmentación no es completamente la adecuada y esto se ve reflejado en los valores de validación del F-Measure.



Adicionalmente, se obtuvo el número de noticias segmentadas por Fuzzy C-Means en las distintas categorías de noticias, la siguiente tabla y gráfica representan la información:

| Categoria | Número de noticias |
|---|---|
| Sport | 409 |
| Politics | 441 |
| Entertainment | 315 |
| Tech | 977 |
| Business | 83 |

# VI. Conclusiones

- **La precisión del algoritmo Fuzzy C-Means se ve impactado por el tamaño del vector de palabras que se recibe como entrada**. En nuestro caso si el universo de palabras era completo las noticias se categorizaban en todos los grupos homogéneamente por lo que su F-Measure disminuía.
- Los parámetros que más afectaron el F-Measure fueron el **MinGain y el Fuzzyness**, esto debido a que son los que definen que tantos elementos se agrupan y definen el nivel de aceptación de cambios en las iteraciones.
- El **desempeño general del clasificador fue menor** en comparación con RN (Accuracy 49%) y AG (F-Measure 55%-75%).
- **Existen términos que proporcionan la misma información en distintos clusters**, por lo que como vimos un cluster puede contener varias categorías, lo que disminuye el F-Measure del FCM.
- Nuevamente, al no aplicar **Stop Words** sobre el vector de palabras, el clasificador obtuvo un menor desempeño.
- Se mantiene la tendencia de que la categoría **Sports** ha sido fácilmente identificada tanto por RN, AG y LD, gracias a su vocabulario especializado en términos concretos de deportes.

# VII. Bibliografía

[1] Charu C. Aggarwal, ChengXiang Zhai, “Mining Text Data”. Springer 2012.

[2] CRISP-DM, SPSS Inc. 2000.

[3] Thaung Thaung Win. Mon, Lin. "Document clustering by fuzzy c-mean algorithm," 2010 2nd International Conference on Advanced Computer Control, 2010, pp. 239-242, doi: 10.1109/ICACC.2010.5487022.

[4] W. Wang, C. Wang, X. Cui and A. Wang, "Fuzzy C-Means Text Clustering with Supervised Feature Selection," 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008, pp. 57-61, doi: 10.1109/FSKD.2008.161.

[5] Torra, Vicenç. (2015). On the selection of m for Fuzzy c-Means. 10.2991/ifsa-eusflat-15.2015.224.

[6] Cano, A. Cano, E. Análisis de los algoritmos de agrupamiento borroso para detectar asimetría de información.