UNIVERSIDAD SANTO TOMÁS

PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

SECCIONAL TUNJA

VIGILADA MINEDUCACIÓN - SNIES 1732

Acreditación Institucional Internacional

OTORGADA POR EL IAC CINDA ACUERDO 55 DEL 8 DE MAYO–VIGENCIA 5 AÑOS

ACREDITACIÓN INSTITUCIONAL DE ALTA CALIDAD MULTICAMPUS

Res. MEN No. 04456 del 29 de enero de 2016

Vigencia por seis años

QS STARS RATED FOR EXCELLENCE

ISO 9001 icontec SC4289–1

IQNet CERTIFIED MANAGEMENT SYSTEM

UNIVERSIDAD SANTO TOMÁS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA
SECCIONAL TUNJA

VIGILADA MINEDUCACIÓN - SNIES 1732

Sentiment Analysis in Twitter

Autor: *Carlos David Páez Ferreira* – ***Fecha:*** *07 de mayo de 2021*

CONTENIDO

1. Metadatos del escrito.
2. Corpus usado.
3. Técnica de aprendizaje.
4. Porcentaje de acierto obtenido.
5. Conclusiones de los investigadores

¡Siempre hacia lo alto!

Metadatos del escrito

Nombre: Happy parents’ tweets: An exploration of Italian Twitter data using sentiment analysis

Año: 2019

Autores: Letizia Mencarini, Delia Irazú Hernández-Farías, Mirko Lai, Viviana Patti, Emilio Sulis and Daniele Vignoli

Corpus Usado

Para hacer el análisis de los datos en italiano, los investigadores hicieron uso de un dataset llamado Twita-2014, el cual comprende un total de 259’893.081 tweets en italiano, de los cuales 4’766.342 han sido geotaggeados entre las 110 provincias italianas y tomando en cuenta la correlación que existía con la edad.

Corpus Usado

Se filtró un grupo de datos de Twita-2014 para seleccionar una submuestra donde los usuarios hablan sobre los temas de interes (padres felices). El filtro se aplica por medio de 11 hashtags, inflexión (diminutivos, singulares y plurales), palabras claves del *Vocabolario di base della lingua italiana (VdB)*.

Corpus

Con lo anterior se logró obtener un conjunto de datos de 3.9 millones de tweets, pero al eliminar tweets “ruidosos” se mantuvieron 2.8 millones de tweets.

Corpus de Twitter “estándar de oro”: Para crear un corpus de oro con una anotación semántica sobre la paternidad, desarrollaron un esquema de anotaciones de varias capas. Pasaron una muestra aleatoria de 6000 tweets por CrowdFlower y se obtuvieron 1508 opiniones etiquetadas.

Técnica de aprendizaje

Seleccionaron los tweets de interés, luego calcularon el sentimiento mediante una técnica de machine learning. Entrenaron con el modelo Support Vector Machine binario y luego hicieron uso de un modelo de “bag-of-words” con características como puntuación, marcas, longitud de tweets, hashtags, menciones, emojis e interjecciones.

El modelo entrenado automáticamente, distinguió que de los 2.8 millones de tweets, 1’083.741 (37%), hablan sobre los temas clasificados

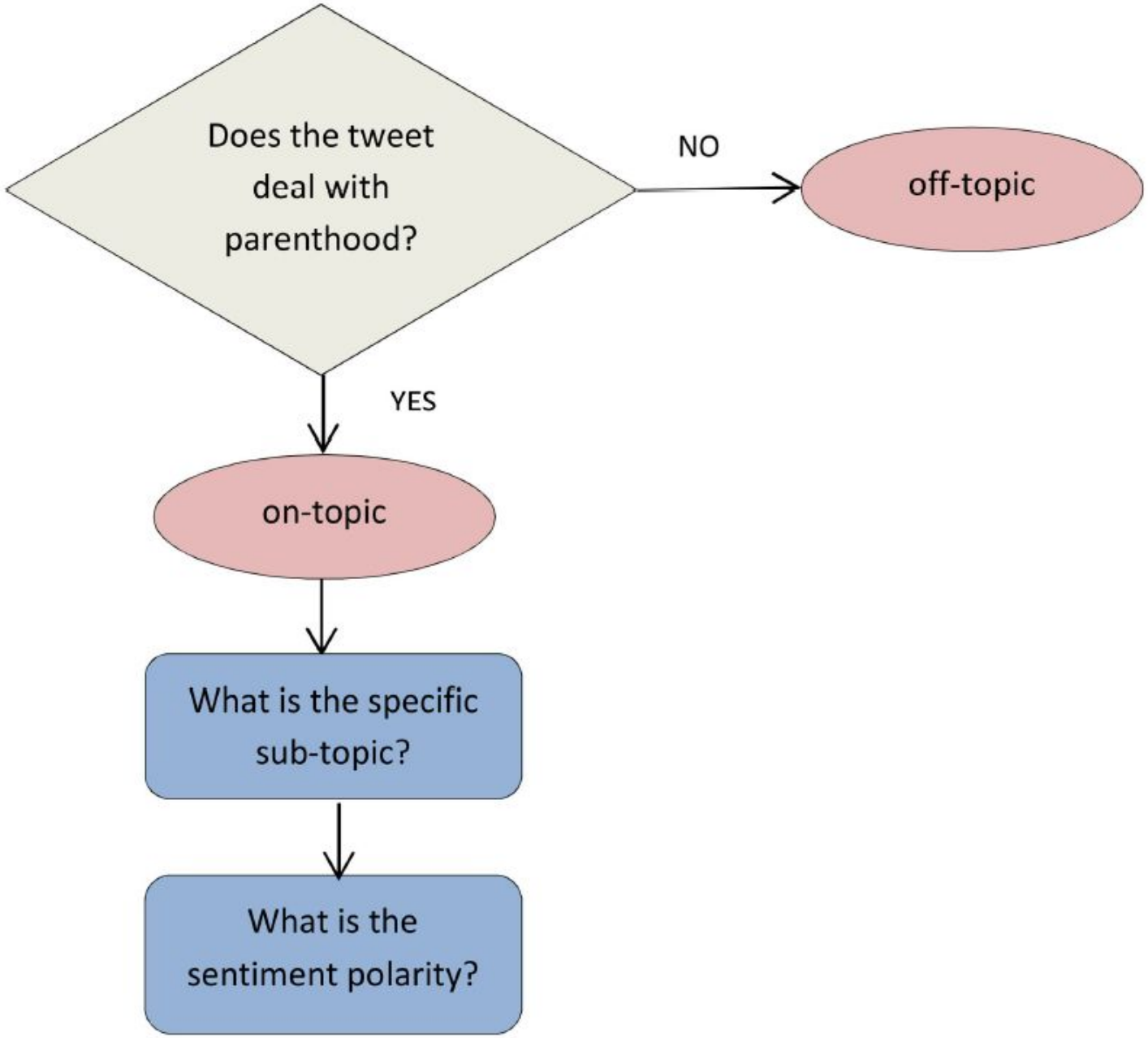


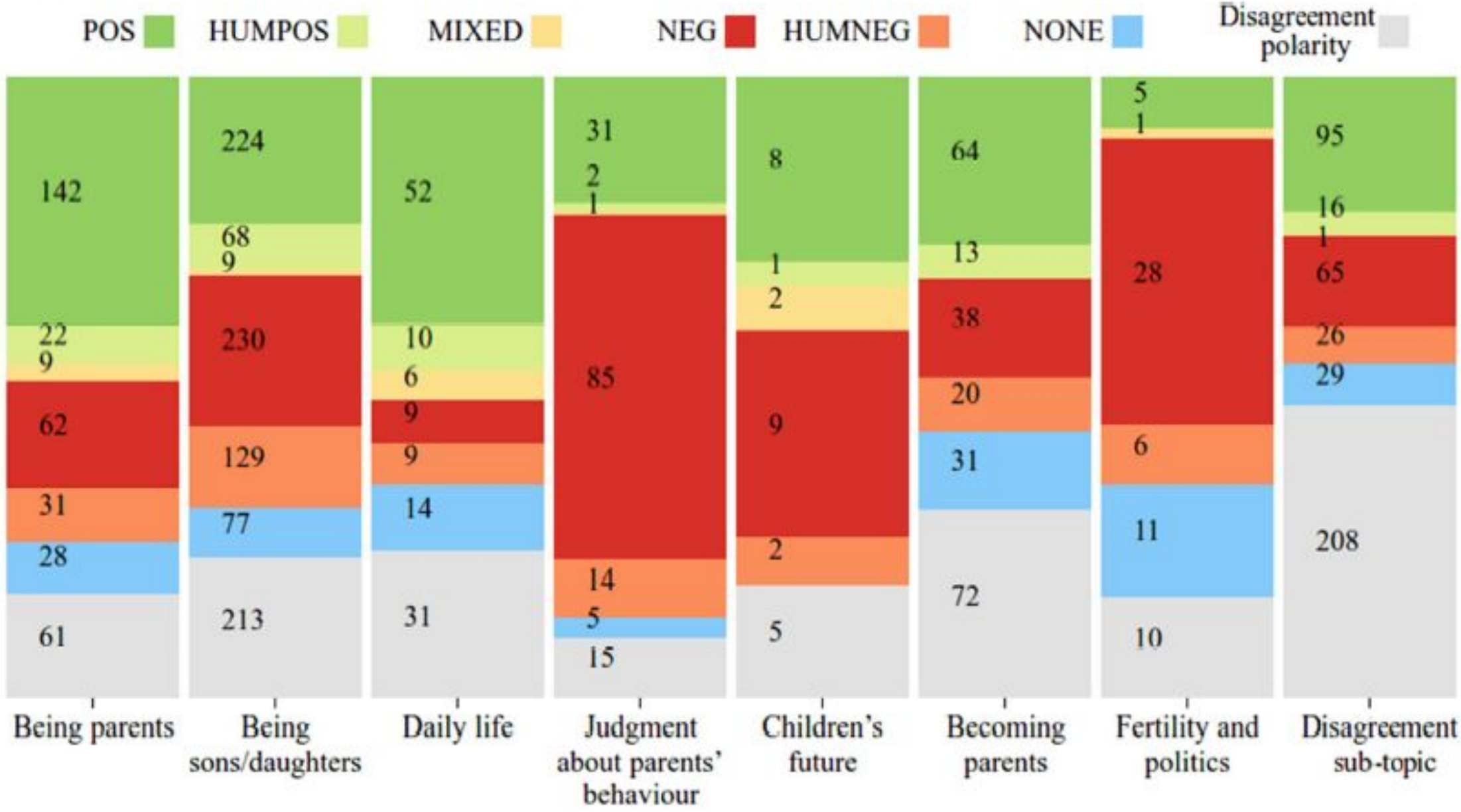
Table 1: Distribution of gold standard messages about parenthood, by polarity

Polarity	Num	%
Positive	526	34.9
Positive humour	116	7.7
Mixed	28	1.8
Negative humour	211	14.0
Negative	461	30.6
None	166	11.0
Total	1,508	100

Table 2: Distribution of gold standard messages about parenthood, by subtopic

Label	Num	%
Being sons/daughters	737	48.9
Being parents	294	19.5
Becoming parents	166	11.0
Judgment about parents' behaviour	138	9.2
Daily life	100	6.6
Fertility and politics	51	3.4
Children's future	22	1.4
Total	1,508	100

Figure 2: Label distribution in our gold corpus and disagreement



Porcentaje de acierto obtenido.

El desempeño del clasificador fue evaluado por la “F-measure”, el cual provee información sobre la precisión, teniendo en cuenta la relación entre precisión y recuperación, entonces, al hacer 5 veces una validación cruzada, se obtuvo un valor de F-measure de 0,7496

Table 4: Distribution of sentiment labels annotated by IRADABE

Class	Tweets	Percentage (%)
Positive	109,272	10.1
Negative	538,127	49.7
Mixed	391,522	36.1
None	44,820	4.1
TOT	1,083,741	100.0

Conclusiones

Es posible que para mejorar el algoritmo clasificador, los datos de edad, sexo y numeros de hijos de los usuarios, hubieran estado dentro del dataset. Pero, por cuestiones de política de datos, Twitter no proporciona los 2 primeros, y el otro es casi imposible de saberlo por medio del internet, aunque estos datos demograficos se pueden obtener de manera manual por medio de encuestas.

Referencias

Mencarini, L., Hernández-Farías, D., Lai, M., Patti, V., Sulis, E., & Vignoli, D. (2019). Happy parents' tweets: An exploration of Italian Twitter data using sentiment analysis. Demographic Research, 40, 693-724. Retrieved May 7, 2021, from https://www.jstor.org/stable/26727014

UNIVERSIDAD SANTO TOMÁS

PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

SECCIONAL TUNJA

VIGILADA MINEDUCACIÓN - SNIES 1732

¡Siempre hacia lo alto!

USTATUNJA.EDU.CO

@santotomastunja