

Universidad “Meyen de San Andrés”

Facultad de Ciencias Puras y Naturales



CURSO DE TEMPORADA

UN ANALISIS DE SENTIMIENTO USANDO R

Universitario(a): Ramiro Brandon Mamani Quisbert

Carrera: Informatica

Docente: Lic. Menfy Morales Rios

Paralelo: A

Fecha: 03 de octubre de 2025

La Paz - Bolivia

1. INTRODUCCIÓN.....	2
1.1 Objetivos del Análisis.....	2
2. METODOLOGÍA.....	2
2.1 Herramientas y Librerías Utilizadas	2
2.2 Fuente de Datos.....	2
3. PROCESO DE ANÁLISIS.....	2
3.1 Análisis de Enlaces Compartidos.....	2
3.2 Procesamiento de Emojis.....	3
3.3 Limpieza y Preprocesamiento del Texto.....	4
3.4 Análisis de Frecuencias.....	6
3.5 Análisis de Sentimientos con Lexicon NRC.....	6
4. RESULTADOS.....	6
4.1 Dominios Web Más Compartidos.....	6
4.2 Palabras Más Frecuentes.....	7
4.3 Distribución de Emociones	8
5. RECOMENDACIONES.....	8
Para Análisis Futuros.....	8
Mejoras Técnicas.....	8
6. CÓDIGO FUENTE.....	8
Estructura del Código.....	8
7. REFERENCIAS.....	9

3. PROCESO DE ANÁLISIS

3.1 Análisis de Enlaces Compartidos

Objetivo: Identificar los dominios web más compartidos en el grupo.

Proceso:

1. Filtrado de mensajes que contienen URLs mediante expresiones regulares
2. Extracción del dominio de cada URL
3. Cálculo de frecuencias y porcentajes
4. Visualización en gráfico de barras horizontal

Hallazgos clave: Los dominios más compartidos reflejan los intereses académicos y de comunicación del grupo (documentos, recursos educativos, plataformas de videoconferencia, etc.).

3.2 Procesamiento de Emojis

Desafío: Los emojis son elementos complejos en el análisis de texto que requieren tratamiento especial.

Solución implementada:

1. Extracción de todos los emojis mediante regex Unicode avanzado que captura:
 - Caras y emociones (U+1F600-U+1F64F)
 - Símbolos y pictogramas (U+1F300-U+1F5FF)
 - Transporte y mapas (U+1F680-U+1F6FF)
 - Emojis modernos y banderas
 - Tonos de piel y secuencias complejas (ZWJ)
1. Creación de un **diccionario personalizado** con más de 200 emojis clasificados en tres categorías:
 - **Positivos:**

```
# Positivos
"😊" = "cara sonriendo", "✅" = "marca de correcto", "✋" = "mano de victoria",
"✨" = "brillo", "💡" = "brillo", "❤️" = "corazon rojo", "❤️🔥" = "corazon en llamas",
"❤️💊" = "corazon vendado", "🌟" = "estrella brillante", "🌸" = "flor de cerezo",
"♣️" = "trebol de la suerte", "🍫" = "chocolate", "🍬" = "caramelo",
"📺" = "caja de regalo", "💣" = "fuegos artificiales", "🌐" = "globo",
"🎉" = "confeti de fiesta", "👏" = "confeti de celebracion", "🏆" = "trofeo",
"👍" = "pulgar arriba", "👍🏠" = "pulgar arriba piel clara", "👏" = "aplausos",
"👏🏠" = "aplausos piel clara", "👏🏠" = "aplausos piel morena", "💪" = "brazo musculoso",
"⚡" = "estrellas de energia", "🔥" = "fuego", "💙" = "corazon azul",
"💜" = "corazon morado", "💜" = "corazon negro", "😬" = "cara con billetes",
"😄" = "cara con estrellas en los ojos", "😄" = "cara sonriendo con lagrima",
"😄" = "cara de fiesta", "😄" = "cara de fiesta", "😄" = "cara de ternura suplicante",
"🤗" = "personas abrazandose", "😭" = "cara derritiendose", "👊" = "cara saludando militarmente",
"😄" = "cara sonriendo", "😄" = "cara sonriendo con ojos grandes", "😄" = "cara llorando de risa",
"😄" = "cara sonriente grande", "😄" = "cara sonriente con ojos de felicidad",
```

- **Negativos:**


```
# Negativos
"☠" = "calavera", "💀" = "calavera", "⚠" = "senal de advertencia", "⚠" = "senal de advertencia",
"✖" = "marca de incorrecto", "!" = "signo de exclamacion", "💔" = "corazon roto",
"👹" = "duende japones", "👹" = "ogro japones", "👻" = "fantasma", "💀" = "calavera",
"💩" = "pila de caca", "🏴‍☠️" = "bandera pirata", "🐉" = "dragon",
"🐍" = "serpiente", "👤" = "cara con simbolos de insultos",
"😬" = "cara con la mano sobre la boca", "😡" = "cara de enojo",
"😭" = "cara llorando", "😓" = "cara de perseverancia", "😓" = "cara triste con sudor",
"😮" = "cara con la boca abierta", "😫" = "cara de dolor agotada", "😴" = "cara de sueno",
"😞" = "cara de agotamiento", "😟" = "cara de disgusto",
"😱" = "cara llorando a gritos", "😨" = "cara de miedo", "😨" = "cara de asombro",
"😳" = "cara avergonzada", "🌀" = "cara con espirales en los ojos",
"😾" = "gato llorando", "😠" = "cara de desaprobacion", "😌" = "cara de alivio con sudor",
"😞" = "cara pensativa triste", "😞" = "cara confundida", "😞" = "cara de sufrimiento",
"😞" = "cara de decepcion", "🙄" = "cara estornudando", "😏" = "cara con ceja levantada",
"😬" = "cara de payaso", "😬" = "cara jadeando", "😬" = "cara con los ojos cubiertos",
"😬" = "cara sin boca", "👤" = "labio mordido"
```

○ **Neutros:**

```
# Neutros / Objetos / Lugares
"🕒" = "cronometro", "☁" = "nube", "☺" = "taza de cafe", "👉" = "dedo indice arriba",
"💎" = "diamante", "✂" = "espadas cruzadas", "⚙" = "engranaje",
"⚡" = "rayo", "⚽" = "balon de futbol", "🛢" = "bomba de gasolina", "🇧🇴" = "bandera de Bolivia",
"✈" = "avion", "👉" = "mano levantada piel morena", "👉" = "mano escribiendo",
"✍" = "lapiz", "✳" = "asterisco", "❓" = "signo de interrogacion",
"➡" = "flecha a la derecha", "⬇" = "flecha hacia abajo", "🌌" = "cielo nocturno galactico",
"🌍" = "planeta tierra europa africa", "🌎" = "planeta tierra americas",
"🌐" = "globo terraqueo", "🌑" = "luna nueva", "🌡" = "termometro", "🎃" = "calabaza",
"🎓" = "birrete de graduacion", "🎙" = "microfono de estudio", "🎞" = "carrete de pelicula",
"🎫" = "boleto", "🎫" = "boleto", "🎙" = "microfono", "🎥" = "camara de cine",
"🎨" = "paleta de artista", "🎫" = "entrada de evento", "🎮" = "control de videojuego",
"🎯" = "diana de tiro", "🎲" = "dados", "🏀" = "balon de baloncesto", "🏐" = "balon de voleibol",
"🏔" = "montana", "🏢" = "edificio de oficinas", "🏫" = "escuela", "🐙" = "pulpo",
"🐧" = "pinguino", "🐯" = "cara de tigre", "👁" = "ojos", "👁" = "ojo en burbuja de dialogo",
"👉" = "dedo indice arriba", "👉" = "dedo indice arriba piel morena", "👇" = "dedo indice abajo"
```

1. Reemplazo de emojis por sus descripciones textuales en español para facilitar el análisis de sentimientos

```
reemplazar_emojis_fijo <- function(textos, dic) {
  for(e in names(dic)) {
    # Agregar espacio antes y después del reemplazo
    textos <- stri_replace_all_fixed(textos, e, paste0(" ", dic[[e]], " "), vectorize_all = FALSE)
  }
  # Quitar espacios al inicio y final y reducir múltiples espacios a uno solo
  textos <- stri_trim_both(textos)
  textos <- stri_replace_all_regex(textos, "\\s+", " ")
  return(textos)
}

df$text <- sapply(df$text, reemplazar_emojis_fijo, dic = diccionario_emoji)
```

3.3 Limpieza y Preprocesamiento del Texto

Pasos aplicados:

1. Normalización básica:

- Conversión a minúsculas
- Eliminación de URLs
- Eliminación de menciones (@usuario)
- Eliminación de puntuación
- Eliminación de números (con excepción de expresiones como "x2" → "yo también")

```
mensajes <- df %>%
  mutate(text_clean = str_to_lower(text),
         text_clean = str_replace_all(text_clean, "http\\S+|www\\S+", ""),
         text_clean = str_replace_all(text_clean, "@\\w+", ""),
         text_clean = str_replace_all(text_clean, "[^\\w\\s]", ""),
         text_clean = str_replace_all(text_clean, "x[0-9]+", "yo tambi
                                     en"),
         text_clean = str_replace_all(text_clean, "[0-9]+", ""),
         text_clean = str_replace_all(text_clean, "ok", "esta bien")) %>%
  filter(text_clean != "")
```

1. Tratamiento de palabras cortas:

- Identificación de expresiones comunes de 1-2 letras
- Reemplazo por categorías semánticas:
 - "sí", "si", "se", "v", "gg" → "afirmación"
 - "no" → "negación"
 - "xd", "gg" → "risa"
 - "hi" → "saludo"

```
dicc_cortas <- tibble::tibble(
  original = c("si", "se", "sí", "no", "xd", "ya", "v", "gg", "hi", "zi", "zy"),
  reemplazo = c("afirmacion", "afirmacion", "afirmacion", "negacion", "risa",
               "afirmacion", "afirmacion", "risa", "saludo", "afirmacion", "afirmacion")
)

mensajes <- mensajes %>%
  left_join(dicc_cortas, by = c("text_clean" = "original")) %>%
  mutate(text_clean = ifelse(!is.na(reemplazo), reemplazo, text_clean)) %>%
  filter(nchar(str_squish(text_clean)) > 2) %>%
  select(-reemplazo)
```

1. Tokenización y filtrado:

- Separación del texto en palabras individuales
- Eliminación de stopwords en español (artículos, preposiciones, etc.)
- Filtrado de tokens de menos de 2 caracteres

```

stop_words <- data.frame(word = stopwords::stopwords("es"))

tokens_clean <- mensajes %>%
  unnest_tokens(word, text_clean) %>%      # tokenizar
  anti_join(stop_words, by = "word") %>%   # eliminar stopwords
  mutate(word = tolower(word),             # pasar a minúsculas
         word = gsub("[^a-záéíóúñ0-9 ]", " ", word), # símbolos a espacio
         word = str_squish(word)) %>%      # eliminar espacios extra
  filter(nchar(word) > 1)                 # eliminar tokens muy cortos
# -----

```

3.4 Análisis de Frecuencias

Se generó un análisis del **Top 10 de palabras más frecuentes** después de la limpieza, presentando:

- Frecuencia absoluta
- Porcentaje respecto al total de palabras
- Visualización en gráfico de barras

3.5 Análisis de Sentimientos con Lexicon NRC

Metodología:

Se utilizó el lexicon **NRC (National Research Council)** que clasifica palabras en 8 emociones básicas:

1. **Ira** (Anger)
2. **Anticipación** (Anticipation)
3. **Disgusto** (Disgust)
4. **Miedo** (Fear)
5. **Alegría** (Joy)
6. **Tristeza** (Sadness)
7. **Sorpresa** (Surprise)
8. **Confianza** (Trust)

Proceso:

1. Cruce de palabras tokenizadas con el lexicon NRC
2. Conteo de frecuencias por categoría emocional
3. Cálculo de porcentajes
4. Visualización con código de colores específico para cada emoción

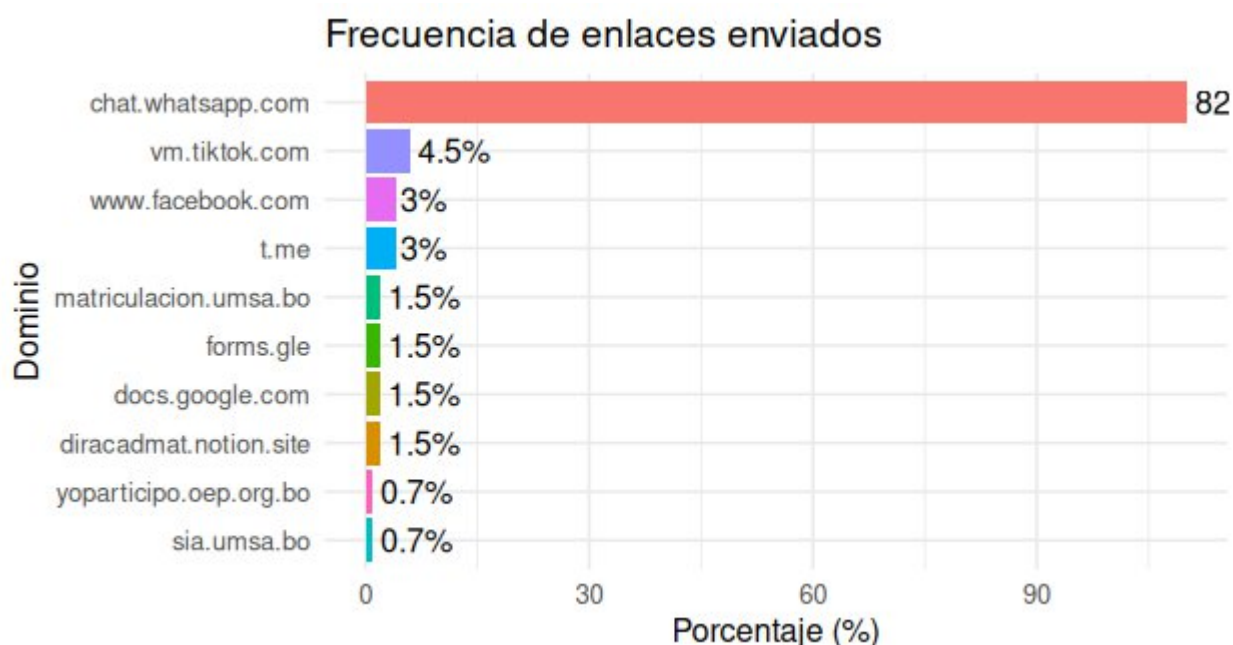

```
emociones_basicas <- c("anger", "anticipation", "disgust", "fear",
                        "joy", "sadness", "surprise", "trust")

tokens_nrc <- tokens_clean %>%
  inner_join(get_sentiments("nrc"), by = c("word" = "word")) %>%
  filter(sentiment %in% emociones_basicas)
```

4. RESULTADOS

4.1 Dominios Web Más Compartidos

Los enlaces compartidos reflejan un uso mixto entre herramientas académicas, plataformas de comunicación y recursos educativos digitales.



Intrepretacion:

chat.whatsapp.com (82%): Este es el hallazgo más significativo. Indica que la abrumadora mayoría de los enlaces enviados en tu grupo de Telegram son invitaciones a un grupo de **whatsapp**.

El grupo de Telegram, en lugar de ser el principal centro de comunicación, está sirviendo principalmente como un "puente" o "portal" para dirigir a los usuarios a otra plataforma de mensajería (WhatsApp).

Matriculacion.umsa.bo (1.5%) y sia.umsa.bo (0.7%): Estos son enlaces **institucionales** directos de la universidad (UMSA). Su baja frecuencia sugiere que los enlaces para trámites oficiales o información académica vital no son el contenido principal que se comparte o reenvía.

forms.gle (1.5%) y docs.google.com (1.5%): Estos son enlaces a **documentos o formularios**, probablemente para encuestas, material de estudio o registro. Aunque son relevantes, su baja proporción confirma que el enfoque del grupo no está en el intercambio constante de material.

diracadmat.notion.site (1.5%): Esto podría ser un enlace útil a una **página de recursos o coordinación** (como Notion), lo cual es positivo, pero marginal en el total.

3. Presencia de Redes Sociales (Contenido Colateral)

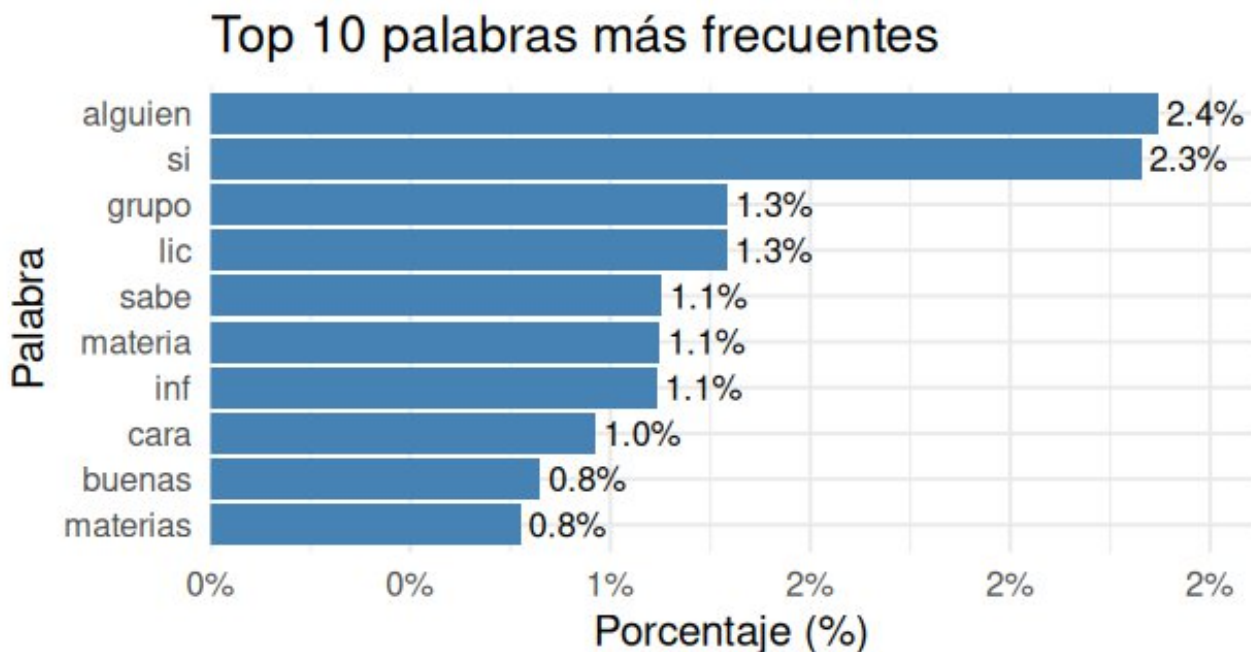
vm.tiktok.com (4.5%) y www.facebook.com (3%): Estos enlaces indican que el grupo se utiliza ocasionalmente para compartir contenido de redes sociales. Esto puede ser relevante para la comunidad (anuncios de la universidad en Facebook, o contenido de ocio/estudio en TikTok), pero **no es el foco principal** ni genera la mayor interacción.

t.me (3%): Curiosamente, estos son enlaces a **otros grupos, canales o mensajes dentro de Telegram**. Esto refuerza que los usuarios envían enlaces, pero incluso dentro de Telegram, son reenvíos a otros chats, no necesariamente contenido nuevo.

4.2 Palabras Más Frecuentes

El vocabulario predominante está relacionado con:

- Términos académicos
- Expresiones de afirmación y acuerdo
- Vocabulario técnico de informática
- Expresiones coloquiales del grupo



Interpretación:

Este análisis de las palabras más comunes proporciona una visión directa del tipo de interacción y las intenciones que predominan en el grupo de Telegram. Confirma que el chat está muy orientado a la coordinación, la consulta y la búsqueda de información, con un fuerte sentido de comunidad.

1. Dominio de la Consulta y la Búsqueda de Ayuda

Las dos palabras más frecuentes, con el doble de ocurrencia que las siguientes, revelan el propósito principal del chat:

- **"alguien" (2.4%):** Este término es el marcador de la **búsqueda de información o ayuda**. Frases como "¿**Alguien** sabe...?", "¿**Alguien** tiene el PDF?" o "¿**Alguien** está en el otro grupo?" son muy comunes.

Implicación: El grupo es visto como un **recurso de apoyo colectivo**. Los estudiantes recurren al grupo para obtener respuestas rápidas de sus pares.

- **"si" (2.3%):** Esto representa la **respuesta afirmativa** a las consultas de "alguien". También puede ser parte de condicionales ("...**si** el licenciado dice...").

Implicación: Hay un alto nivel de **interacción y respuesta** entre los miembros. Cuando alguien pregunta, recibe una respuesta.

2. Enfoque en el Curso, Docentes y Contenido

El segundo bloque de palabras se centra en los elementos directos del entorno académico:

- **"grupo" (1.3%) y "lic" (1.3%):**

"grupo" probablemente se refiere al curso, al chat o, como vimos antes, al otro **grupo de WhatsApp**. Refleja un enfoque en la **coordinación de la comunidad**.

"lic" (abreviatura de licenciado, o docente) indica que el **profesor** es un tema de conversación central (preguntas sobre clases, tareas, o la opinión del docente).

- **"sabe" (1.1%), "materia" (1.1%), "inf" (1.1%):**

"sabe" complementa a "alguien" ("¿Alguien **sabe**...?").

"materia" e **"inf"** (probablemente abreviatura de "Informática" o "información") confirman que el contenido de estudio y la asignatura son el **tema de fondo** de las conversaciones.

3. Fórmulas de Saludo y Coordinación

Las palabras restantes indican las normas de cortesía y la naturaleza de las interacciones:

- **"cara" (1.0%):** En el contexto boliviano de la UMSA, esta palabra probablemente se refiere a las **clases presenciales** o a las actividades "cara a cara", contrastando con la modalidad virtual.
- **"buenas" (0.8%):** Usado en saludos ("Buenos días/tardes/noches"), lo que es común en cualquier chat.

- **"materias" (0.8%):** La forma plural de "materia" sugiere que no solo se habla del curso actual, sino también de **otros cursos** o de la carga académica en general.

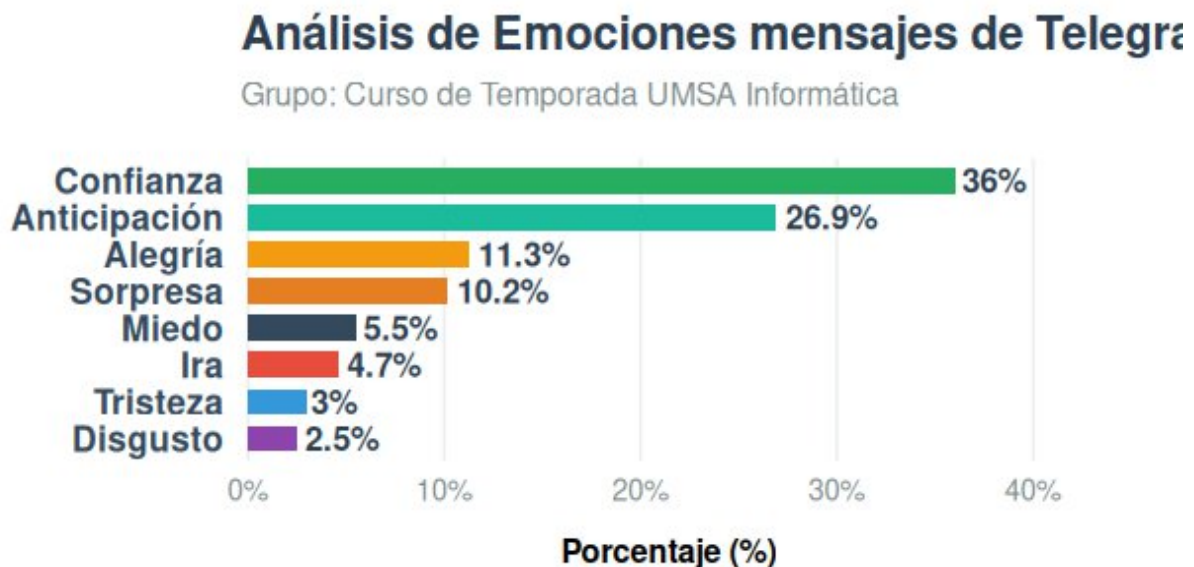
Conclusión

El lenguaje del grupo está dominado por la **logística y la colaboración entre pares**:

1. **Colaboración Activa:** El chat está muy vivo y es **funcional** para su propósito de **apoyo mutuo**. Las personas preguntan ("**alguien**") y las personas responden ("**si**").
2. **Enfoque Claro:** La conversación está centrada en los temas correctos para un grupo de curso: el **docente** ("lic"), el **contenido** ("materia", "inf") y la **modalidad** ("cara").
3. **Refuerzo del Problema de Canal:** Si bien la mayoría de las palabras son útiles, la alta frecuencia de "grupo" podría seguir siendo un indicador de que los estudiantes están constantemente **preguntando o refiriéndose al otro grupo de WhatsApp**.

4.3 Distribución de Emociones

El análisis de sentimientos reveló la distribución porcentual de 8 emociones básicas presentes en los mensajes del grupo. Los resultados se presentan en un gráfico de barras horizontales con código de colores:



Interpretación del análisis de emociones:

Este análisis de sentimientos ofrece una **excelente perspectiva** del estado emocional general de los miembros de tu grupo de Telegram del Curso de Temporada UMSA Informática. El panorama es abrumadoramente **positivo y proactivo**.

1. El Sentimiento Predominante: Positivo y Orientado al Futuro

Las dos emociones más frecuentes, **Confianza** y **Anticipación**, son cruciales y suman casi el **63%** del total de emociones detectadas.

- **Confianza (36%)**: Esta es la emoción dominante. Indica que los participantes se sienten **seguros** respecto a la información que reciben, la **capacidad de la organización** o su **propio desempeño** y éxito en la materia. Es un fuerte indicador de **seguridad y estabilidad** en el grupo.
- **Anticipación (26.9%)**: Este alto porcentaje refleja un **fuerte compromiso e interés**. Los estudiantes están activamente **esperando** el inicio de clases, la próxima actividad o la publicación de notas. Es una emoción que impulsa a la acción y demuestra **motivación**.

2. La Atmósfera General: Alegría y Dinamismo

La siguiente capa de emociones refuerza el tono positivo y añade un elemento de interacción agradable.

- **Alegría (11.3%)**: El ambiente en el chat no es puramente formal; un 11.3% de mensajes alegres indica que hay un **tono optimista y satisfecho**. Los miembros disfrutaban de las interacciones o están celebrando pequeños logros/noticias positivas.
- **Sorpresa (10.2%)**: Este porcentaje alto puede ser ambivalente, pero en el contexto de alta **Confianza** y **Alegría**, es probable que se deba a:
 - **Anuncios inesperados** (ej. "¡Sorpresa! El examen se retrasa" o "Recursos extra liberados").
 - **Interacciones espontáneas** o reacciones a las respuestas de otros.

3. Las Preocupaciones: Minoritarias pero Presentes

Las emociones negativas (Miedo, Ira, Tristeza y Disgusto) representan solo el **15.7%** del total, lo cual es un porcentaje bajo y muy positivo.

- **Miedo (5.5%)**: Es la emoción negativa más alta. Suele estar relacionada con la **incertidumbre académica** (dificultad del material, exámenes o la carga de trabajo).
 - *Sugerencia*: Dado que existe **mucha Confianza**, el miedo puede ser a **fallar** a pesar de su esfuerzo. Reforzar el apoyo o aclarar los criterios de evaluación podría ayudar a mitigar esta preocupación.
- **Ira (4.7%), Tristeza (3%), Disgusto (2.5%)**: Estos bajos porcentajes sugieren que no hay problemas de **conflicto, frustración o descontento** generalizado. Lo más probable es que estos mensajes sean reacciones aisladas a trámites, problemas técnicos o frustraciones personales, y no un sentimiento dirigido al curso o a la administración del grupo.

Conclusión

El análisis de emociones te da una señal muy clara:

El grupo tiene una salud emocional excelente: Los estudiantes se sienten **confiados** y **motivados**, lo que se refleja en una alta **Anticipación**.



5. RECOMENDACIONES

Para Análisis Futuros

1. **Análisis temporal:** Estudiar cómo evolucionan las emociones a lo largo del semestre
2. **Análisis por usuario:** Identificar patrones individuales de comunicación
3. **Modelado de topics:** Usar LDA (Latent Dirichlet Allocation) para identificar temas de conversación
4. **Análisis de redes:** Estudiar patrones de interacción entre miembros

Mejoras Técnicas

1. Implementar un lexicon de sentimientos específico para español
2. Ampliar el diccionario de emojis con actualizaciones periódicas
3. Considerar bigramas y trigramas para capturar mejor el contexto
4. Aplicar técnicas de machine learning para clasificación de sentimientos

6. CÓDIGO FUENTE

El análisis completo fue desarrollado en R y está disponible con todas las transformaciones, visualizaciones y procesamiento de datos documentado.

Estructura del Código

1. Carga de librerías y datos
2. Análisis de URLs compartidas
3. Procesamiento de emojis con diccionario personalizado
4. Limpieza y normalización de texto
5. Tokenización y eliminación de stopwords
6. Análisis de frecuencias
7. Análisis de sentimientos con NRC
8. Generación de visualizaciones

7. REFERENCIAS

- **NRC Emotion Lexicon:** Mohammad, S. M., & Turney, P. D. (2013). "Crowdsourcing a Word-Emotion Association Lexicon"
- **tidytext:** Silge, J., & Robinson, D. (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R"
- **ggplot2:** Wickham, H. (2016). "ggplot2: Elegant Graphics for Data Analysis"