

# Tabla de Contenido

<b>1. Introducción</b>	<b>1</b>
1.1. Contexto de Trabajo . . . . .	1
1.2. Antecedentes Generales . . . . .	2
1.2.1. Chile y el consumo de alcohol . . . . .	2
1.2.2. Relación entre alcohol y suicidio . . . . .	3
1.2.3. Importancia de las redes sociales e Internet . . . . .	4
1.2.4. Web Intelligence Centre . . . . .	4
1.3. Descripción del Proyecto y Justificación . . . . .	5
1.4. Objetivos . . . . .	6
1.4.1. Objetivo general . . . . .	6
1.4.2. Objetivos específicos . . . . .	6
1.5. Hipótesis de investigación . . . . .	7
1.6. Metodología . . . . .	7
1.7. Resultados esperados . . . . .	8
1.8. Estructura del informe . . . . .	8
<b>2. Marco Teórico</b>	<b>10</b>
2.1. World Wide Web . . . . .	10
2.1.1. Web 2.0 . . . . .	10
2.2. Proceso Knowledge Discovery in Databases (KDD) . . . . .	11
2.2.1. Data Mining . . . . .	12
2.3. Web Mining . . . . .	12
2.3.1. Definición de Web Intelligence . . . . .	12
2.3.2. Categorías . . . . .	13
2.3.3. Web Opinion Mining . . . . .	14
2.4. Técnicas de Minería de Datos . . . . .	14
2.5. Text Mining . . . . .	15
2.6. Machine Learning (ML) . . . . .	15
2.7. Algoritmos Supervisados y no Supervisados . . . . .	16
2.8. Topping Modeling . . . . .	16
2.9. Redes Sociales . . . . .	17
2.9.1. Microblogging . . . . .	17
2.9.2. Twitter . . . . .	17
2.9.3. Interfaz de Programación de Aplicaciones (API) . . . . .	19
2.10. Extracción de información . . . . .	22
2.10.1. Crawling . . . . .	22

2.10.2. Preprocesamiento de datos . . . . .	22
2.11. Validación Cruzada o <i>Cross Validation</i> . . . . .	25
2.11.1. K-Fold Cross-validation . . . . .	25
2.12. Kappa de Fleiss . . . . .	26
2.13. Evaluación de rendimiento . . . . .	26
<b>3. Consumo de alcohol</b>	<b>28</b>
3.1. Factores que afectan el consumo de alcohol y daños relacionados con el alcohol . .	28
3.1.1. Edad . . . . .	28
3.1.2. Género . . . . .	29
3.1.3. Factores de riesgo familiar . . . . .	29
3.1.4. Estatus socio-económico . . . . .	30
3.1.5. Control y regulación de alcohol . . . . .	30
3.2. Los daños relacionados al alcohol . . . . .	32
3.2.1. Trastornos por consumo de alcohol . . . . .	32
3.3. Alcohol Use Identification Test (AUDIT) . . . . .	33
<b>4. Diseño</b>	<b>35</b>
4.1. Requerimientos . . . . .	35
4.1.1. Variables originales de la Encuesta Nacional de Drogas . . . . .	35
4.1.2. Segmentación original de la Encuesta Nacional de Drogas . . . . .	38
4.2. Indicadores finales utilizados para el estudio en <i>Twitter</i> . . . . .	38
4.3. Descripción de los datos utilizados . . . . .	39
4.3.1. Datos disponibles . . . . .	40
4.3.2. Estructura de los datos . . . . .	40
4.3.3. Etiquetado de <i>Tweets</i> . . . . .	42
4.3.4. Etiquetado de usuarios . . . . .	42
4.3.5. Selección de <i>Keywords</i> . . . . .	43
4.4. Diseño de la aplicación . . . . .	44
4.4.1. Tratamiento de texto . . . . .	44
4.4.2. Cálculo de la polaridad de <i>tweets</i> . . . . .	45
4.4.3. Cálculo de la edad en usuarios . . . . .	45
4.4.4. Atributos para el usuario . . . . .	45
<b>5. Implementación</b>	<b>48</b>
5.1. Herramientas utilizadas . . . . .	48
5.2. Selección de palabras claves o <i>Keywords</i> . . . . .	53
5.3. Etiquetado . . . . .	55
5.3.1. Etiquetado de <i>Tweets</i> . . . . .	55
5.3.2. Etiquetado de usuarios . . . . .	59
5.4. Mantenimiento de datos . . . . .	60
<b>6. Resultados</b>	<b>68</b>
6.1. Palabras claves o <i>Keywords</i> . . . . .	68
6.2. Recolección de datos . . . . .	69
6.3. Etiquetado . . . . .	72
6.3.1. Etiquetado de <i>tweets</i> . . . . .	72

6.3.2.	Etiquetado de usuarios . . . . .	74
6.4.	Evaluación de algoritmos . . . . .	79
6.4.1.	Detección de consumo en <i>tweets</i> . . . . .	80
6.4.2.	Detección de políticas en <i>tweets</i> . . . . .	81
6.4.3.	Consumo de alcohol en usuarios . . . . .	82
6.4.4.	Consumo de riesgo de alcohol en usuarios . . . . .	83
6.5.	Métricas . . . . .	86
6.5.1.	Prevalencia . . . . .	86
6.5.2.	Consumo de riesgo . . . . .	87
6.5.3.	Frecuencia de consumo . . . . .	88
6.5.4.	Polaridad . . . . .	89
6.5.5.	Polaridad de políticas . . . . .	90
6.5.6.	Palabras más utilizadas . . . . .	90
<b>7.</b>	<b>Trabajo futuro y conclusiones</b>	<b>100</b>
7.1.	Conclusiones generales . . . . .	100
7.1.1.	Ética . . . . .	101
7.2.	Trabajo futuro . . . . .	102
7.2.1.	Mejoras a los modelos de clasificación de <i>Tweets</i> . . . . .	102
7.2.2.	Mejoras a los modelos de clasificación de usuarios . . . . .	103
	<b>Bibliografía</b>	<b>107</b>

# Índice de tablas

2.1. Matriz de confusión. . . . .	27
3.1. Dominios y contenidos de los items del Test AUDIT . . . . .	34
4.1. Datos del usuario disponibles en <i>Twitter</i> y útiles para el estudio. . . . .	41
4.2. Datos del <i>tweet</i> disponibles en <i>Twitter</i> y útiles para el estudio. . . . .	41
6.1. Tabla de términos utilizados. . . . .	70
6.2. Número de cuentas nuevas de usuarios chilenos creadas por año y número acumulado de cuentas de usuarios chilenos. . . . .	71
6.3. Número de <i>tweets</i> por año que contienen las keywords seleccionadas. . . . .	71
6.4. Heterogeneidad en las etiquetas . . . . .	74
6.5. Medidas de acuerdo en el primer etiquetado. . . . .	74
6.6. Porcentaje de prevalencia de la muestra. . . . .	76
6.7. Escala AUDIT (OMS). . . . .	79
6.8. Escala AUDIT Validación Chile. . . . .	80
6.9. Rendimiento de <i>Naive Bayes</i> para la detección de consumo en el primer etiquetado . . . . .	80
6.10. Rendimiento de <i>SVM</i> para la detección de consumo en el segundo etiquetado . . . . .	80
6.11. Rendimiento de <i>SVM</i> para la detección de políticas en el primer etiquetado . . . . .	81
6.12. Rendimiento de <i>SVM</i> para la detección de políticas en el segundo etiquetado . . . . .	81
6.13. Rendimiento de <i>SVM</i> para la detección de consumo de alcohol en usuarios . . . . .	82
6.14. Influencia de variables en el consumo de alcohol . . . . .	83
6.15. Descripción de las clases usadas en modelo de consumo de riesgo . . . . .	84
6.16. Rendimiento de <i>SVM</i> para la detección de consumo de riesgo en usuarios. . . . .	84
6.17. Influencia de variables en el consumo de riesgo de alcohol . . . . .	85
6.18. Frecuencia de palabras en <i>tweets</i> de consumo. . . . .	92
6.19. Consumo de bebidas alcohólicas en Chile en el año 2015. . . . .	94
6.20. Frecuencia de palabras en <i>tweets</i> de políticas. . . . .	94
7.1. <i>Tweets</i> en los que la información está contenida en el <i>hashtag</i> . . . . .	103
7.2. Tabla de Emojis asociados a alcohol. . . . .	105
7.3. Descripción de usuarios en donde manifiestan que consumen alcohol . . . . .	106

# Índice de figuras

1.1. Carga de AVISA (años) atribuible a Factores de Riesgo según género, Chile 2007. . . . .	3
2.1. Diagrama del proceso KDD. . . . .	12
2.2. Técnicas de minería de datos. . . . .	15
2.3. Dos aplicaciones comunicadas utilizando una API . . . . .	19
2.4. Streaming API . . . . .	20
2.5. Rest API . . . . .	20
2.6. Arquitectura de un Web Crawler . . . . .	22
2.7. Cross-validation. Procedimiento de <i>three-fold cross validation</i> . . . . .	26
5.1. Ejemplo de formato JSON. . . . .	51
5.2. Ejemplo de archivo Arff . . . . .	54
5.3. Bienvenida del Sitio Web de la encuesta . . . . .	60
5.4. Consentimiento informado del Sitio Web de la encuesta . . . . .	61
5.5. Primeras tres preguntas del Sitio Web de la encuesta . . . . .	62
5.6. Preguntas AUDIT en el Sitio Web de la encuesta . . . . .	63
5.7. Modelo E-R de alcoholdb . . . . .	64
5.8. Modelo E-R de tagging . . . . .	65
5.9. Modelo E-R de usertrace . . . . .	66
5.10. Modelo de relaciones utilizando ArangoDB . . . . .	67
6.1. Porcentaje de cuentas públicas y privadas en <i>Twitter</i> . . . . .	69
6.2. Número de cuentas creadas total acumulado por año . . . . .	72
6.3. Número de <i>tweets</i> relacionados con alcohol por año. . . . .	73
6.4. Distribución de edad de la muestra . . . . .	75
6.5. Distribución de la prevalencia de la muestra . . . . .	76
6.6. Distribución de lo encuestados según sexo . . . . .	77
6.7. Distribución de la muestra según puntaje AUDIT . . . . .	78
6.8. Prevalencia de alcohol por mes separado según sexo . . . . .	79
6.9. Evolución de la prevalencia de consumo de alcohol en el último mes . . . . .	87
6.10. Evolución de la prevalencia de consumo de alcohol en el último mes . . . . .	88
6.11. Comparación de las dos curvas de evolución de la prevalencia de consumo de alcohol en el último mes. . . . .	89
6.12. Evolución del consumo de riesgo y dependencia sobre el total de la muestra por cada año . . . . .	90
6.13. Evolución del consumo de riesgo y dependencia sobre el total de la población general	91

6.14. Comparación de las dos curvas de evolución de la prevalencia de consumo de riego de alcohol. . . . .	93
6.15. Promedio de <i>Tweets</i> de Consumo por Usuario por Año . . . . .	95
6.16. Evolución del promedio de días de consumo de alcohol en el último mes. . . . .	96
6.17. Evolución del promedio de la polaridad en los <i>tweets</i> de alcohol. . . . .	97
6.18. Evolución del promedio de la polaridad de los usuarios. . . . .	98
6.19. Evolución del promedio de la polaridad de los <i>Tweets</i> de políticas. . . . .	99