Capítulo 1

Procesamiento del Lenguaje Natural

El volumen de datos textuales digitalizados que se genera cada día es enorme (por ejemplo, la web, redes sociales, registros médicos, libros digitalizados). Por lo tanto, también crece la necesidad de traducir, analizar y gestionar esta avalancha de palabras y texto.

El procesamiento del lenguaje natural (PLN) es el campo que se encarga de diseñar métodos y algoritmos que toman como entrada o producen como salida datos de **lenguaje natural** no estructurado [Goldberg, 2017]. El PLN se centra en el diseño y análisis de algoritmos computacionales y representaciones para procesar el lenguaje humano [Eisenstein, 2018].

Una tarea común de PLN es el Reconocimiento de Entidades Nombradas (NER, por sus siglas en inglés). Por ejemplo:

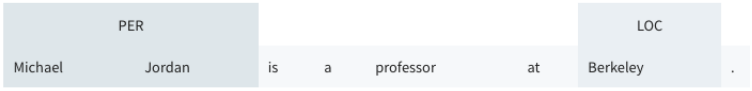


Figura 1.1: Reconocimiento de Entidades Nombradas

El lenguaje humano es altamente ambiguo, como en las frases: Çomí pizza con amigos", Çomí pizza con aceitunas.º Çomí pizza con un tenedor". Además, el lenguaje está en constante cambio y evolución, como ocurre con los hashtags en Twitter.

1.1. PLN y Lingüística Computacional

El procesamiento del lenguaje natural (PLN) desarrolla métodos para resolver problemas prácticos relacionados con el lenguaje [Johnson, 2014].

Algunos ejemplos son:

- Reconocimiento automático del habla.
- Traducción automática.
- Extracción de información de documentos.

La lingüística computacional (LC) estudia los procesos computacionales subyacentes al lenguaje (humano).

- ¿Cómo comprendemos el lenguaje?
- ¿Cómo producimos el lenguaje?
- ¿Cómo aprendemos el lenguaje?

El PLN y la LC utilizan métodos y modelos similares.

Aunque existe una superposición sustancial, hay una diferencia importante en el enfoque. La LC se centra en la lingüística respaldada por métodos computacionales (similar a la biología computacional o la astronomía computacional). En lingüística, el lenguaje es el objeto de estudio. El PLN se centra en resolver tareas bien definidas relacionadas con el lenguaje humano (como la traducción, la respuesta a consultas, las conversaciones). Si bien los conocimientos lingüísticos fundamentales pueden ser cruciales para realizar estas tareas, el éxito se mide en función de si y cómo se logra el objetivo (según una métrica de evaluación) [Eisenstein, 2018].

El procesamiento del lenguaje natural y la lingüística computacional están estrechamente relacionados y se superponen en muchos aspectos. Ambos campos utilizan métodos y modelos similares para abordar problemas relacionados con el lenguaje humano. Sin embargo, la diferencia principal radica en el enfoque: la lingüística computacional se centra en la lingüística respaldada por métodos computacionales, mientras que el procesamiento del lenguaje natural se centra en resolver tareas prácticas relacionadas con el lenguaje. Ambos campos son fundamentales para comprender y aprovechar el poder del lenguaje humano en la era digital.

1.2. Niveles de descripción lingüística

El campo de la **descripción lingüística** abarca diferentes niveles:

- **Fonética y fonología:** estudio de los sonidos del habla.
- **Morfología:** estudio de la estructura de las palabras.

- **Sintaxis:** estudio de la estructura de las oraciones.
- **Semántica:** estudio del significado de las palabras y oraciones.
- **Pragmática:** estudio del uso del lenguaje en el contexto.

El PLN puede abordar tareas en cada uno de estos niveles, pero a menudo se enfoca en niveles más altos de representación y comprensión.

1.2.1. Fonética

La fonética es la rama de la lingüística que se ocupa del estudio de los sonidos del lenguaje. Examina los órganos utilizados en la producción de sonidos, como la boca, la lengua, la garganta, la nariz, los labios y el paladar. Los sonidos del lenguaje se dividen en vocales y consonantes. Las vocales se producen con poca restricción del flujo de aire desde los pulmones, mientras que las consonantes implican alguna restricción o cierre en el tracto vocal [Johnson, 2014, Fromkin et al., 2018]. Además, el Alfabeto Fonético Internacional (AFI) proporciona una notación alfabética para representar los sonidos fonéticos.

1.2.2. Fonología

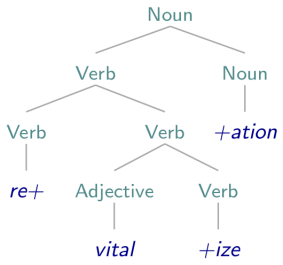
La fonología se centra en el estudio de cómo los sonidos del habla forman patrones y construyen significado. Los fonemas son las unidades básicas de sonido que diferencian el significado de las palabras. Por ejemplo, en inglés, la "pz la "b"son fonemas distintos porque cambian el significado de las palabras en las que se encuentran. La fonología también examina las variaciones en la pronunciación de los sonidos en diferentes contextos y dialectos [Fromkin et al., 2018].

1.2.3. Morfología

La morfología se ocupa del estudio de la estructura interna de las palabras. Los morfemas son las unidades mínimas de significado que componen las palabras. Por ejemplo, en la palabra "deshacer", los morfemas son "des-", "hacerz er". La morfología también se interesa por los procesos de formación de palabras, como la derivación, donde se agregan prefijos o sufijos a una palabra existente para formar una nueva palabra con un significado diferente [Johnson, 2014].

- Morphology studies the structure of words (e.g.,re+structur+ing, un+remark+able) [Johnson, 2014]
- Morpheme: The linguistic term for the most elemental unit of grammatical form [Fromkin et al., 2018]. Example morphology= morph + ology (the science of).

- Derivational morphology: process of forming a new word from an existing word, often by adding a prefix or suffix
- Derivational morphology exhibits a hierarchical structure. Example: re+vital+ize+ation

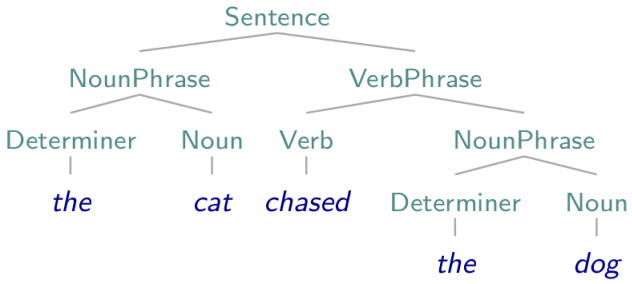


- The suffix usually determines the syntactic category (part-of-speech) of the derived word.

1.2.4. Sintaxis

La sintaxis es el estudio de cómo las palabras se combinan para formar frases y oraciones gramaticales. Examina las reglas y estructuras que determinan la organización de las palabras en una oración y cómo influyen en el significado. La sintaxis también se ocupa de la relación entre las palabras y las funciones que desempeñan dentro de una oración. Por ejemplo, en la oración .[El] perro persigue al gato", .[el] perro.[es] el sujeto, "persigue.[es] el verbo y .[a]l gato.[es] el complemento directo [Johnson, 2014].

- Syntax studies the ways words combine to form phrases and sentences [Johnson, 2014]



- Syntactic parsing helps identify **who did what to whom**, a key step in understanding a sentence.

1.2.5. Semántica

La semántica es el estudio del significado de las palabras, frases y oraciones, examinando cómo se construye e interpreta este significado en el contexto del lenguaje. Además, la semántica se interesa por los roles semánticos, que indican la función de cada entidad en una oración. Por ejemplo, en la oración .El niño cortó la cuerda con una navaja", .el niño.es el agente, "la cuerda.es el tema y üna navaja.es el instrumento [Johnson, 2014].

La semántica se enfoca en el significado de las palabras, frases y oraciones. Estudia cómo se construye e interpreta este significado en el contexto del lenguaje. Además, dentro de la semántica, se analizan los roles semánticos, los cuales indican la función que desempeña cada entidad en una oración. Por ejemplo, en la oración .El niño cortó la cuerda con una navaja", se identifican distintos roles semánticos: .el niñoçomo el agente, "la cuerdaçomo el tema y üna navajaçomo el instrumento utilizado [Johnson, 2014].

En resumen:

- La semántica estudia el significado de las palabras, frases y oraciones [Johnson, 2014].
- Dentro de la semántica, se analizan los roles semánticos, que indican el papel desempeñado por cada entidad en una oración.
- Algunos ejemplos de roles semánticos son: **agente** (la entidad que realiza la acción), **tema** (la entidad involucrada en la acción) y **instrumento** (otra entidad utilizada por el agente para llevar a cabo la acción).
- En la oración .El niño cortó la cuerda con una navaja", se puede identificar el agente como **el niño**, el tema como **la cuerda** y el instrumento como **una navaja**.
- Además de los roles semánticos, la semántica también abarca las relaciones léxicas, que son las relaciones entre diferentes palabras [Yule, 2016].
- Algunos ejemplos de relaciones léxicas incluyen la sinonimia (conceal/hide), la antonimia (shallow/deep) y la hiponimia (perro/animal).

1.2.6. Pragmática

La pragmática se centra en cómo el contexto influye en la interpretación y el significado de las expresiones lingüísticas. Examina cómo se utilizan las expresiones lingüísticas en situaciones reales y cómo los hablantes interpretan el significado implícito. Por ejemplo, la oración "Hace frío aquí"puede interpretarse como una sugerencia implícita de cerrar las ventanas [Fromkin et al., 2018].

1.3. Procesamiento del Lenguaje Natural y Aprendizaje Automático

Aunque los seres humanos somos grandes usuarios del lenguaje, también somos muy malos para comprender y describir formalmente las reglas que rigen el lenguaje.

Entender y producir lenguaje utilizando computadoras es altamente desafiante. Los métodos más conocidos para lidiar con datos de lenguaje se basan en el aprendizaje automático supervisado.

El aprendizaje automático supervisado consiste en intentar inferir patrones y regularidades a partir de un conjunto de pares de entrada y salida preanotados (también conocido como conjunto de datos de entrenamiento).

Conjunto de Datos de Entrenamiento: Datos de NER CoNLL-2003 Cada línea contiene un token, una etiqueta de parte de la oración, una etiqueta de sintagma y una etiqueta de entidad nombrada.

```
U.N.          NNP  I-NP  I-ORG
official      NN   I-NP  O
Ekeus         NNP  I-NP  I-PER
heads         VBZ  I-VP  O
for           IN   I-PP  O
Baghdad       NNP  I-NP  I-LOC
.             .    O     O
```

[1]Fuente: `https://www.clips.uantwerpen.be/conll2003/ner/`

1.4. Desafíos del Lenguaje

Existen tres propiedades desafiantes del lenguaje: la discreción, la composicionalidad y la dispersión.

Discreción: no podemos inferir la relación entre dos palabras a partir de las letras que las componen (por ejemplo, hamburguesa y pizza).

Composicionalidad: el significado de una oración va más allá del significado individual de sus palabras.

Dispersión: la forma en que las palabras (símbolos discretos) pueden combinarse para formar significados es prácticamente infinita.

1.5. Ejemplo de tareas NLP

Clasificación de temas La clasificación de temas es una tarea de Procesamiento del Lenguaje Natural (PLN) en la cual se asigna a un documento una de varias categorías, como deportes, política, cotilleos o economía. Las palabras presentes en los documentos brindan pistas importantes sobre su tema. Sin

embargo, redactar reglas para esta tarea es un desafío debido a la complejidad del lenguaje. La anotación de datos, en la cual los lectores clasifican los documentos por temas, puede ayudar a generar conjuntos de datos de entrenamiento para algoritmos de aprendizaje automático supervisado. Estos algoritmos aprenden patrones de uso de palabras que facilitan la categorización de los documentos.

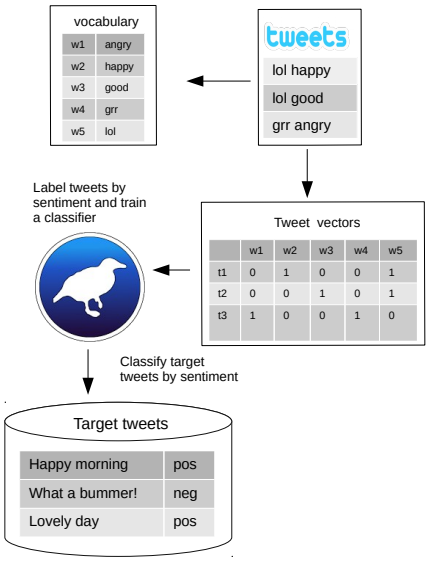
- Clasificar un documento en una de las cuatro categorías: Deportes, Política, Cotilleos y Economía.
- Las palabras en los documentos proporcionan indicios muy sólidos.
- ¿Qué palabras brindan qué indicios?
- Elaborar reglas para esta tarea resulta bastante desafiante.
- No obstante, los lectores pueden categorizar fácilmente varios documentos según su tema (anotación de datos).
- Un algoritmo de aprendizaje automático supervisado puede identificar los patrones de uso de palabras que ayudan a categorizar los documentos.

Análisis de Sentimiento El análisis de sentimientos se refiere a la aplicación de técnicas de Procesamiento del Lenguaje Natural (PLN) para identificar y extraer información subjetiva de conjuntos de datos textuales. Un desafío común en el análisis de sentimientos es la clasificación de la polaridad a nivel de mensaje (MPC), donde las frases se clasifican automáticamente en categorías positivas, negativas o neutrales. Las soluciones más avanzadas utilizan modelos de aprendizaje automático supervisado entrenados con ejemplos anotados manualmente.

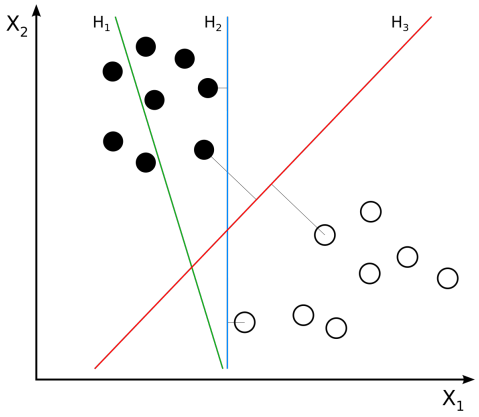
En este tipo de clasificación, es habitual emplear el aprendizaje supervisado, siendo las Máquinas de Vectores de Soporte (SVM) una opción popular. El objetivo de las SVM es encontrar un hiperplano que separe las clases con el margen máximo, logrando la mejor separación entre las clases positivas, negativas y neutrales [Eisenstein, 2018].

- Aplicación de técnicas de **PLN** para identificar y extraer información subjetiva de conjuntos de datos textuales.
- Clasificación automática de frases en las categorías **positiva**, **negativa** o **neutral**.

- Las soluciones más avanzadas emplean modelos de aprendizaje automático **supervisado**, entrenados con ejemplos **anotados manualmente** [Mohammad et al., 2013].



- Idea: Encontrar un hiperplano que separe las clases con el margen máximo (mayor separación).



- H_3 separa las clases con el margen máximo.

1.5.1. Lingüística y Procesamiento del Lenguaje Natural (PNL)

El conocimiento de las estructuras lingüísticas es fundamental para el diseño de características y el análisis de errores en el Procesamiento del Lenguaje

Natural (PNL). Los enfoques de aprendizaje automático en PNL se basan en características que describen y generalizan las instancias de uso del lenguaje. El conocimiento lingüístico orienta la selección y el diseño de estas características, ayudando al algoritmo de aprendizaje automático a encontrar correlaciones entre el uso del lenguaje y las etiquetas objetivo [Bender, 2013].

- El conocimiento de las estructuras lingüísticas es importante para el diseño de características y el análisis de errores en PNL [Bender, 2013].
- Los enfoques de aprendizaje automático en PNL requieren características que puedan describir y generalizar el uso del lenguaje.
- El objetivo es guiar al algoritmo de aprendizaje automático para encontrar correlaciones entre el uso del lenguaje y el conjunto de etiquetas objetivo.
- El conocimiento sobre las estructuras lingüísticas puede influir en el diseño de características para los enfoques de aprendizaje automático en PNL.

El PNL plantea diversos desafíos, como los costos de anotación, las variaciones de dominio y la necesidad de actualizaciones continuas. La anotación manual requiere mucho trabajo y tiempo. Las variaciones de dominio implican aprender patrones diferentes para diferentes corpus de texto. Los modelos entrenados en un dominio pueden no funcionar bien en otro. Además, los modelos de PNL pueden volverse obsoletos a medida que el uso del lenguaje evoluciona con el tiempo.

1.6. Desafíos en el Procesamiento del Lenguaje Natural (PNL)

- **Costos de Anotación**: la anotación manual es **laboriosa** y **consume mucho tiempo**.
- **Variaciones de Dominio**: el patrón que queremos aprender puede variar de un corpus a otro (por ejemplo, deportes, política).
- ¡Un modelo entrenado con datos anotados de un dominio no necesariamente funcionará en otro!
- Los modelos entrenados pueden quedar desactualizados con el tiempo (por ejemplo, nuevos hashtags).

Variación de Dominio en el Análisis de Sentimiento

1. Para mí, la cola era bastante **pequeña** y solo tuve que esperar unos 20 minutos, ¡pero valió la pena! :D @raynwise

2. Extraña espacialidad en Stuttgart. La habitación del hotel es tan **pequeña** que apenas puedo moverme, pero los alrededores son inhumanamente vastos y largos bajo construcción.

Superando los costos de anotación de datos Supervisión Distant:

- Etiquetar automáticamente datos no etiquetados (**API de Twitter**) utilizando un método heurístico.
- **Enfoque de Anotación de Emoticonos (EAA)**: los tweets con emoticonos positivos **:)** o negativos **:(** se etiquetan según la polaridad indicada por el emoticono [Read, 2005].
- El emoticono se **elimina** del contenido.
- Este enfoque también se ha ampliado utilizando hashtags como #anger y emojis.
- No es trivial encontrar técnicas de supervisión distante para todo tipo de problemas de PNL.

Crowdsourcing

- Confiar en servicios como **Amazon Mechanical Turk** o **Crowdflower** para solicitar a la **multitud** que anote datos.
- Esto puede resultar costoso.
- Es difícil garantizar la calidad de las anotaciones.

1.7. Estudio de caso: Clasificación de sentimientos en tweets

- En 2013, el taller de Evaluación Semántica (SemEval) organizó la tarea de .ªnálisis de sentimientos en Twitter"[Nakov et al., 2013].
- La tarea se dividió en dos sub-tareas: el nivel de expresión y el nivel del mensaje.
- Nivel de expresión: se centró en determinar la polaridad del sentimiento de un mensaje según una entidad marcada dentro de su contenido.
- Nivel del mensaje: se debía determinar la polaridad según el mensaje en general.
- Los organizadores lanzaron conjuntos de datos de entrenamiento y prueba para ambas tareas [Nakov et al., 2013].

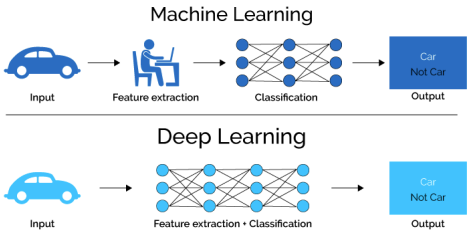
El sistema NRC

- El equipo que logró el mejor rendimiento en ambas tareas, entre 44 equipos, fue el equipo *NRC-Canada* [Mohammad et al., 2013].
- El equipo propuso un enfoque supervisado utilizando un clasificador SVM lineal con las siguientes características hechas a mano para representar los tweets:

  1. N-gramas de palabras.
  2. N-gramas de caracteres.
  3. Etiquetas de partes del discurso.
  4. Agrupaciones de palabras entrenadas con el método de agrupamiento de Brown [Brown et al., 1992].
  5. El número de palabras alargadas (palabras con un carácter repetido más de dos veces).
  6. El número de palabras con todas las letras en mayúscula.
  7. La presencia de emoticonos positivos o negativos.
  8. El número de negaciones individuales.
  9. El número de secuencias contiguas de puntos, signos de interrogación y signos de exclamación.
  10. Características derivadas de lexicones de polaridad [Mohammad et al., 2013]. Dos de estos lexicones se generaron utilizando el método PMI a partir de tweets anotados con hashtags y emoticonos.

1.8. Ingeniería de características y Aprendizaje Profundo

- Hasta 2014, la mayoría de los sistemas de PNL de última generación se basaban en ingeniería de características + modelos de aprendizaje automático superficiales (por ejemplo, SVM, HMM).
- Diseñar las características de un sistema de PNL ganador requiere mucho conocimiento específico del dominio.
- El sistema NRC se construyó antes de que el aprendizaje profundo se hiciera popular en PNL.
- Por otro lado, los sistemas de Aprendizaje Profundo se basan en redes neuronales para aprender automáticamente buenas representaciones.



Ingeniería de características y Aprendizaje Profundo

- El Aprendizaje Profundo proporciona resultados de última generación en la mayoría de las tareas de PNL.
- Grandes cantidades de datos de entrenamiento y máquinas GPU multicore más rápidas son clave en el éxito del aprendizaje profundo.
- Las **redes neuronales** y las **incrustaciones de palabras** desempeñan un papel fundamental en los modelos modernos de PNL.

Aprendizaje Profundo y Conceptos Lingüísticos

- Si los modelos de aprendizaje profundo pueden aprender representaciones automáticamente, ¿siguen siendo útiles los conceptos lingüísticos (por ejemplo, sintaxis, morfología)?
- Algunos defensores del aprendizaje profundo argumentan que estas propiedades lingüísticas inferidas y diseñadas manualmente no son necesarias, y que la red neuronal aprenderá estas representaciones intermedias (o equivalentes o mejores) por sí misma [Goldberg, 2016].
- Aún no hay un consenso definitivo al respecto.
- Goldberg cree que muchos de estos conceptos lingüísticos pueden ser inferidos por la red por sí misma si se le proporciona suficiente cantidad de datos.
- Sin embargo, en muchos otros casos no disponemos de suficientes datos de entrenamiento para la tarea que nos interesa, y en estos casos proporcionar a la red los conceptos generales más explícitos puede ser muy valioso.

1.9. Historia

El progreso de la PNL se puede dividir en tres oleadas principales: 1) racionalismo, 2) empirismo y 3) aprendizaje profundo [Deng and Liu, 2018].

1950 - 1990 Racionalismo: se enfocaba en diseñar reglas hechas a mano para incorporar conocimiento y mecanismos de razonamiento en sistemas de PNL inteligentes (por ejemplo, ELIZA para simular a un psicoterapeuta Rogeriano, MARGIE para estructurar información del mundo real en ontologías de conceptos).

1991 - 2009 Empirismo: se caracteriza por la explotación de corpora de datos y modelos de aprendizaje automático y estadísticos (superficiales) (por ejemplo, Naive Bayes, HMMs, modelos de traducción IBM).

2010 - Aprendizaje Profundo: la ingeniería de características (considerada como un cuello de botella) se reemplaza con el aprendizaje de representaciones y/o redes neuronales profundas (por ejemplo, `https://www.deepl.com/translator`). Un artículo muy influyente en esta revolución: [Collobert et al., 2011].

1.10. Conclusiones

En este capítulo, hemos explorado el desafío de entender y producir lenguaje utilizando computadoras. El aprendizaje automático supervisado es una de las principales técnicas utilizadas para abordar este desafío. Además, hemos discutido las propiedades desafiantes del lenguaje, como la discreción, la composicionalidad y la dispersión. Estos aspectos nos muestran la complejidad inherente al procesamiento del lenguaje natural y nos desafían a encontrar soluciones efectivas.

[1] Las fechas son aproximadas.

Bibliografía

[Bender, 2013] Bender, E. M. (2013). Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax. *Synthesis lectures on human language technologies*, 6(3):1–184.

[Brown et al., 1992] Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

[Collobert et al., 2011] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537.

[Deng and Liu, 2018] Deng, L. and Liu, Y. (2018). *Deep Learning in Natural Language Processing*. Springer.

[Eisenstein, 2018] Eisenstein, J. (2018). Natural language processing. Technical report, Georgia Tech.

[Fromkin et al., 2018] Fromkin, V., Rodman, R., and Hyams, N. (2018). *An introduction to language*. Cengage Learning.

[Goldberg, 2016] Goldberg, Y. (2016). A primer on neural network models for natural language processing. *J. Artif. Intell. Res.(JAIR)*, 57:345–420.

[Goldberg, 2017] Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

[Johnson, 2014] Johnson, M. (2014). Introduction to computational linguistics and natural language processing (slides). 2014 Machine Learning Summer School.

[Mohammad et al., 2013] Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*.

[Nakov et al., 2013] Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T. (2013). Semeval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises*, pages 312–320, Atlanta, Georgia, USA. Association for Computational Linguistics.

[Read, 2005] Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, ACLstudent '05, pages 43–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Yule, 2016] Yule, G. (2016). *The study of language*. Cambridge university press.