Academic Year 2023/2024

Social Media Analytics (SMA)

Sentiment Analysis

Part 2

Marco Viviani

University of Milano-Bicocca

Department of Informatics, Systems, and Communication

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

Approaches to Sentiment Analysis

- Existing approaches to (affective computing and) sentiment analysis can be grouped into three main categories:
  - Knowledge-based techniques.
  - Supervised methods.
  - Hybrid approaches.

Knowledge-based Techniques

- **Knowledge-based techniques** are very popular because of their accessibility and economy.

- Text is classified into affect categories based on the presence of fairly unambiguous affect words like ‘happy’, ‘sad’, ‘afraid’, and ‘bored’.

- They are based on the usage of **lexicons** to perform both emotion analysis and sentiment analysis.

Knowledge-based Techniques: Weaknesses

- The major weakness of knowledge-based approaches is **poor recognition of affect** when linguistic rules are involved.
  - For example, while a knowledge base can correctly classify the sentence “today was a happy day” as being happy, it is likely to fail on a sentence like “today wasn’t a happy day at all”.
  - To this end, more sophisticated knowledge-based approaches exploit linguistics rules to distinguish how each specific knowledge base entry is used in text.

- Another limitation of knowledge-based approaches lies in the typicality of their **knowledge representation**.
  - Knowledge is usually strictly defined and does not allow handling different concept nuances, as the inference of semantic and affective features associated with concepts is bounded by the fixed, flat representation.

Supervised Methods

- **Supervised methods** (such as machine and deep learning), have been popular for affect classification of texts and have been used by many researchers.

- By feeding a machine learning algorithm a large training corpus of affectively **annotated texts**, it is possible for the system to not only learn the affective valence of affect keywords (as in the keyword spotting approach), but also to take into account the valence of other arbitrary keywords (like lexical affinity) and word co-occurrence frequencies.

Hybrid Approaches

- **Hybrid approaches** to affective computing and sentiment analysis exploit both knowledge-based techniques and supervised methods to perform tasks such as emotion recognition and polarity detection from text or multimodal data.

- Given lexicons can be used as additional features in supervised approaches to improve the effectiveness of the considered model.

Sentiment Analysis Resources (Lexicons and Datasets)

WordNet

- **WordNet** (Miller 1995) is a lexical database of the English language that groups words into sets of synonyms (synsets), providing short definitions and usage examples.
  - It is a large lexical database of English, developed at Princeton University.
  - https://wordnet.princeton.edu/

- WordNet links words into semantic relations, including synonyms (words that have similar meanings), hypernyms (words that represent more abstract or general concepts), hyponyms (words that represent more specific concepts), and meronyms (words that denote a part of something).

- It has been used as a basis for some sentiment lexicons.

Sentiment Lexicons

- Early development of sentiment lexicons focused on creation of sentiment dictionaries.
- Thet are collections of words or phrases that are annotated with their associated sentiment or emotion.
- Each word or phrase in the dictionary is labeled as positive, negative, or neutral, indicating the sentiment conveyed by that particular linguistic element.
- They can be either manually-generated or automatically-generated.

Manually-generated Sentiment Lexicons

- Stone et al. (1966) present a lexicon called **General Inquirer** (GI) that has been widely used for sentiment analysis.
  - Stone, Philip J., Dexter C. Dunphy, and Marshall S. Smith. "The general inquirer: A computer approach to content analysis." (1966).
  - It has sentiment labels for about 3,600 terms.

- Finn (2011) present a lexicon called **AFINN**.
  - Like General Inquirer, it is also a manually generated lexicon.
  - http://corpustext.com/reference/sentiment_afinn.html

AFFIN

- **Scoring system:** it scores words on a scale from -5 to +5. A negative score indicates a negative sentiment, a positive score indicates a positive sentiment, and a score of 0 represents neutrality.
  - The magnitude of the score indicates the intensity of the sentiment.
- **Limited to unigrams:** it primarily focuses on unigrams, which are single words.
  - It does not explicitly consider the sentiment of multi-word phrases or the context in which words are used.
  - This makes it a relatively simple and straightforward resource.
- **Use in Sentiment Analysis:** it is commonly used in sentiment analysis tasks, especially those where a lightweight and easy-to-implement solution is desired.

Automatically-generated Sentiment Lexicons

- **SentiWordNet**, described first by Esuli and Sebastiani (2006), is a sentiment lexicon which augments WordNet (Miller 1995) with sentiment information.
  - Sebastiani, Fabrizio, and Andrea Esuli. "Sentiwordnet: A publicly available lexical resource for opinion mining." Proceedings of the 5th international conference on language resources and evaluation. European Language Resources Association (ELRA) Genoa, Italy, 2006.
  - https://github.com/aesuli/SentiWordNet
- **SenticNet** introduced by Cambria et al. (2016) has sentiment entries for 30,000 words and multi-word expressions using information propagation to connect various parts of common-sense knowledge representations.
  - Cambria, Erik, et al. "SenticNet." Sentic Computing: a common-sense-based framework for concept-level sentiment analysis (2015): 23-71.
  - https://sentic.net/

SentiWordNet (1)

- The labelling is **fuzzy** and is done by adding three sentiment scores to each synset (a synset refers to a set of synonymous words or phrases that represent a single concept) in the WordNet as follows.

- Every synsets has three scores:
  1. Pos(s): The positive score of synsets
  2. Neg(s): The negative score of synsets
  3. Obj(s): The objective score of synsets

SentiWordNet (2)

- In SentiWordNet, sentiment is associated with the **meaning** of a word rather than the word itself.

- This representation allows a word to have multiple sentiments corresponding to each meaning.

- Because there are three scores, each meaning in itself can be both positive and negative, or neither positive nor negative.

VADER (1)

- **Valence Aware Dictionary and sEntiment Reasoner** is a pre-built sentiment analysis tool designed to analyze the sentiment of text data, such as sentences or paragraphs.

- It was developed by researchers at the Georgia Institute of Technology.
  - Hutto, Clayton, and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." Proceedings of the international AAAI conference on web and social media. Vol. 8. No. 1. 2014.
  - https://github.com/cjhutto/vaderSentiment

VADER (2)

- **Lexicon-based**: a pre-built sentiment lexicon that includes words and phrases annotated with their sentiment polarity (positive, negative, or neutral) and intensity.
- **Scores**: two main outputs: a numerical score representing the overall sentiment polarity of the text and an intensity score.
- **Negations and intensifiers**: it handles such linguistic features that influence sentiment.
- **Rules-based**: the algorithm considers a combination of factors, including individual word polarities, the presence of capitalization and punctuation, and the context of words within a sentence.
- **Social media**: it is particularly well-suited for sentiment analysis in the context of social media and informal text.
  - Presence of slang, emojis, and other informal language elements).
  - Short texts, such as tweets or product reviews, where traditional sentiment analysis methods may face challenges.

Emotion Lexicons

- **Emotion lexicons** are specialized dictionaries or lists of words and phrases annotated with their associated emotional content.

- These lexicons provide information about the emotional or affective meaning of words, helping in the analysis of sentiment, mood, or emotions expressed in written or spoken language.

- Emotion lexicons are commonly used in natural language processing (NLP) and computational linguistics for tasks such as sentiment analysis, emotion detection, and affective computing.

LIWC (1)

- **LIWC**, or Linguistic Inquiry and Word Count, is a text analysis program that analyzes written or spoken language based on linguistic and psychological dimensions.

- Developed by James W. Pennebaker and his colleagues, LIWC provides insights into the psychological and emotional content of text by categorizing words into various linguistic and psychological dimensions.
  - https://www.liwc.app/

LIWC (2)

- **Word categories**: 90 different linguistic and psychological dimensions, including categories such as emotions, cognitive processes, social processes, and more.
- **Customizable dictionaries**: researchers can adapt LIWC to their needs by defining word categories relevant to their study.
- **Applications**: psychology, linguistics, communication studies, and social sciences.
- **Emotion and sentiment analysis**: LIWC includes categories related to emotions and sentiment, making it useful for sentiment analysis and understanding the emotional content of textual data.

WordNet-Affect

- **WordNet-Affect** (Strapparava and Valitutti 2004) like SentiWordNet, is a resource that annotates senses in WordNet with emotions.
  - https://aclanthology.org/L04-1208/

- It consists of 2,874 synsets annotated with affective labels (called a-labels).

- WordNet-Affect was created using a semi-supervised method as follows:
  - A set of core synsets is created.
  - These are synsets whose emotion has been manually labelled in the form of a-labels.
  - These labels are projected to other synsets using WordNet relations.
  - The a-labels are then manually evaluated and corrected, wherever necessary.

EmoLex (NRC Emotion Lexicon)

- **EmoLex**, short for "Emotion Lexicon," is a lexicon or dictionary of words that are annotated with their associated emotional content.
  - https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
- Emotional categories: words are categorized into different emotional categories, such as joy, anger, sadness, fear, disgust, and surprise.
  - Each word in the lexicon is assigned a score or label indicating its association with one or more emotions.
- Valence scores: EmoLex often includes valence scores that represent the degree of positive or negative sentiment associated with each word.
  - This allows for a more nuanced understanding of emotional content.

SenticNet

- **SenticNet** is a publicly available resource for sentiment analysis and concept-level opinion mining: https://sentic.net/
- **Polarity scores**: they indicate the degree of positivity or negativity associated with that concept.
  - Polarity scores range from -1 (very negative) to 1 (very positive).
- **Semantic roles**: these roles provide information about how each concept contributes to the overall sentiment of a sentence or text.
- **Affects**: SenticNet includes information about affects associated with each concept. Affects represent the basic emotional qualities linked to a concept.
- **Multilingual support**: SenticNet is available in multiple languages, making it versatile for sentiment analysis and opinion mining in diverse linguistic contexts.

Annotated Datasets (1)

- **SemEval** is a competition that is run for specific tasks.
  - Sentiment analysis and related tasks have featured since 2013.
  - The datasets for these tasks are released online.
  - https://semeval.github.io/

- Distinct tasks
  - https://aclanthology.org/S17-2088/
  - https://aclanthology.org/2022.semeval-1.180/

Annotated Datasets (2)

- The **Stanford NLP group**
  - The Stanford Natural Language Processing (NLP) Group is a research group within the Department of Computer Science at Stanford University that focuses on advancing the field of natural language processing through research, development, and education.
  - https://nlp.stanford.edu/

- Sentiment Analysis @ Stanford
  - https://nlp.stanford.edu/sentiment/index.html

Annotated Datasets (3)

- The **Amazon Review** Dataset
  - This dataset contains information regarding product information (e.g., color, category, size, and images) and more than 230 million customer reviews from 1996 to 2018. The reviews are labeled based on their positive, negative, and neutral emotional tone.

- The **Yelp Review** Dataset
  - This open-source dataset includes more than 500,000 training samples consisting of consumer reviews, ratings, and recommendations. The polarity score of each sentence is determined, and the keywords requested can be extracted.