Twitter Sentiment Analysis

Milestone 1

Avinash Alapati

Department of Master Data Science Program

Bellevue University

DSC680-T301 Applied Data Science (2243-1)

DSC 680 -PROJECT PROPOSAL MILESTONE 1

Avinash Alapati

TOPIC – TWITTER SENTIMENT ANALYSIS

This project aims to conduct sentiment analysis on a selected product or service, to be determined as the project unfolds. The primary data source for opinions will be Twitter. The overarching objective of the sentiment analysis is to uncover the user perception of the chosen product or service. Extracted opinions will be categorized into positive, neutral, and negative sentiments. Subsequently, an analysis will be conducted on the classified data to determine the percentage distribution across these three sentiment categories within the sampled population.

BUSINESS PROBLEM

Sentiment analysis is the automated extraction of attitudes, opinions, views, and emotions from various sources such as text, speech, tweets, and databases, employing Natural Language Processing (NLP). This process entails categorizing textual opinions into classifications such as "positive," "negative," or "neutral." Additionally, sentiment analysis is commonly known as subjectivity analysis, opinion mining, and appraisal extraction.

The terms opinion, sentiment, view, and belief are often used interchangeably, yet they carry distinct meanings:

- Opinion: A conclusion that may be open to dispute, especially when various experts hold differing viewpoints.
- View: A subjective opinion or perspective.
- Belief: The intentional acceptance and intellectual agreement with a certain idea or concept.
- Sentiment is the expression of one's feelings and opinions.

In contemporary times, sentiment analysis and Natural Language Processing (NLP) hold significant importance. The internet witnesses a massive daily influx of information on social media platforms and blogs, much of which computers struggle to comprehend. Traditionally, processing such vast datasets was challenging, but advancements in computer performance, aligning with Moore's law, and the advent of distributed computing technologies like Hadoop or Apache Spark have made it more manageable. Ongoing research and investment in this field promise a future where computers can derive understanding from text, greatly enhancing data analytics and search engine capabilities.

For certain companies, understanding a customer's perception of their product is invaluable data. Analyzing such information allows companies to identify product issues, discern trends ahead of competitors, enhance communication with their target audience, and gain insightful feedback on the effectiveness of their marketing campaigns. This knowledge serves as valuable input for the development of the next generation of their products.

METHODS

This segment will outline the technical methodology adopted for the project, encompassing a description of the system.

Data Mining

Data, specifically tweets, will be gathered utilizing Twitter's API. The Twitter API, an open-source Python library, provides access to the complete set of Twitter's RESTful API features. This library will be employed to retrieve tweets, and the outcomes will be compared with the parsing method.

Data Cleansing

A tweet encapsulates diverse opinions expressed in various ways by different users. The Twitter dataset employed in this project has already been categorized into two classes: negative and positive polarity, facilitating the ease of observing the impact of different features on sentiment analysis. The raw data, with polarity labels, is susceptible to inconsistency and redundancy. Tweet preprocessing encompasses several steps:

1. Eliminating all URLs (e.g., www.xyz.com), hash tags (e.g., #topic), and targets (@username).
2. Spell correction, handling sequences of repeated characters.
3. Substituting emoticons with their corresponding sentiments.
4. Removing all punctuations, symbols, and numbers.
5. Eliminating stop words.
6. Expanding acronyms using an acronym dictionary.
7. Discarding non-English tweets.

In the subsequent phase of the system, the collected data will undergo cleansing, involving the removal of punctuations and conversion to lowercase. This preparation will prove beneficial in the subsequent stages of the project, particularly in the "Bag of Words" approach, as eliminating lowercase words reduces redundancy in the database storing the words.

Classifying the data

Machine learning techniques necessitate the representation of key features in text or documents for effective processing. These features, considered as feature vectors, play a crucial role in classification tasks. Various examples of features reported in the literature include:

1. Words and Their Frequencies: Unigrams, bigrams, and n-gram models with frequency counts serve as features.
2. Parts of Speech Tags: Parts of speech, such as adjectives, adverbs, and specific groups of verbs and nouns, provide valuable indicators of subjectivity and sentiment. Syntactic dependency patterns can be generated through parsing or dependency trees.
3. Opinion Words and Phrases: Beyond individual words, phrases and idioms conveying sentiments, such as "cost someone an arm and a leg," can be utilized as features.
4. Position of Terms: The location of a term within a text influences its impact on the overall sentiment.
5. Negation: Negation is a crucial but challenging feature to interpret, as its presence can alter the polarity of an opinion.
6. Syntax: Syntactic patterns like collocations are employed as features to learn subjectivity patterns.

This phase is expected to be the most challenging in the project, involving the analysis of individual words or word groups in a tweet and attempting to assign sentiment. This task is complex, especially given the difficulty for a computer to comprehend slang words and sarcasm.

The "Bag of Words" model will employ a database of positive, negative, and neutral words. Each tweet will be dissected into individual words and compared to those in the databases. Matching words will increment or decrement a counter based on assigned weightings. The final counter will then be used to classify the sentiment; for instance, a tweet with predominantly positive words should yield a high counter.

Data Analysis

Once the data is classified, subsequent analysis becomes essential. This analysis may involve straightforward percentages to gauge customer satisfaction, or it could delve into more intricate comparisons. For instance, a more complex analysis might be undertaken to compare customer sentiment regarding two similar products, aiming to uncover correlations between positive sentiment and high sales of those products. This multifaceted examination will provide valuable insights into the relationship between customer perceptions and product performance.

ETHICAL CONSIDERATIONS

User information for all Tweets is maintained in an anonymous manner. Additionally, no demographic details are included for any individual user or group of users sharing the same sentiment. This deliberate exclusion aims to mitigate bias towards any specific user group representing a particular section of society.

CHALLENGES

Following are some of the challenges faced in Sentiment Analysis:

1. Identifying Subjective Parts of Text: Recognizing subjective content in text is challenging as the same word can be subjective in one context and objective in another. For instance:
   - a. "The language of Mr. John was very crude."
   - b. "Crude oil is obtained by extraction from the seabeds." The word "crude" is subjective in the first example but entirely objective in the second.
2. Domain Dependence: Sentences or phrases may have different meanings in various domains. For example, the word "unpredictable" is positive in the context of movies but takes on a negative connotation when referring to a vehicle's steering.
3. Sarcasm Detection: Identifying sarcasm involves recognizing sentences where negative sentiments about a target are expressed using positive words. Example:
   - "Nice perfume. You must shower in it." Though containing only positive words, the sentence conveys a negative sentiment.
4. Thwarted Expressions: Some sentences have specific parts that determine the overall polarity of the document. Example:
   - "This movie should be amazing. It sounds like a great plot, the popular actors, and the supporting cast are talented as well." While a simple bag-of-words approach may classify it as positive sentiment, the ultimate sentiment is negative.

5. Explicit Negation of Sentiment: Sentiment negation can occur in various ways beyond using simple negation words like "no," "not," or "never." Identifying such negations is challenging. Example:
    - "It avoids all suspense and predictability found in Hollywood movies." Here, the words "suspense" and "predictable" carry a negative sentiment, despite the absence of explicit negation words.

REFERENCES

1. Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. LREC, 1320-1326.
2. Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. CS224N Project Report, Stanford.
3. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment Analysis of Twitter Data. In Proceedings of the Workshop on Languages in Social Media (LSM).
4. Barbosa, L., & Feng, J. (2010). Robust Sentiment Detection on Twitter from Biased and Noisy Data. In Proceedings of COLING.
5. Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! In Proceedings of the Fifth International Conference on Weblogs and Social Media.
6. Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a Word-Emotion Association Lexicon. Computational Intelligence, 29(3), 436-465.
7. Davidov, D., Tsur, O., & Rappoport, A. (2010). Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. Proceedings of COLING, 241-249.
8. Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, 2(1-2), 1-135.