Information Fusion 27 (2016) 95–110

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Opinion Mining and Information Fusion: A survey

Jorge A. Balazs, Juan D. Velásquez *

Department of Industrial Engineering, Universidad de Chile, Av. República 701, P.O. Box: 8370439, Santiago, Chile

ARTICLE INFO

Article history:
Received 31 March 2015
Received in revised form 4 June 2015
Accepted 8 June 2015
Available online 17 June 2015

Keywords:
Information Fusion
Survey
Opinion Mining
Sentiment Analysis

ABSTRACT

Interest in Opinion Mining has been growing steadily in the last years, mainly because of its great number of applications and the scientific challenge it poses. Accordingly, the resources and techniques to help tackle the problem are many, and most of the latest work fuses them at some stage of the process. However, this combination is usually executed without following any defined guidelines and overlooking the possibility of replicating and improving it, hence the need for a deeper understanding of the fusion process becomes apparent. Information Fusion is the field charged with researching efficient methods for transforming information from different sources into a single coherent representation, and therefore can be used to guide fusion processes in Opinion Mining. In this paper we present a survey on Information Fusion applied to Opinion Mining. We first define Opinion Mining and describe its most fundamental aspects, later explain Information Fusion and finally review several Opinion Mining studies that rely at some point on the fusion of information.

1. Introduction

With the advent of the Web 2.0 and its continuous growth, the amount of freely available user-generated data has reached an unprecedented volume. Being so massive, it is impossible for humans to make sense of its whole in a reasonable amount of time, which is why there has been a growing interest in the scientific community to create systems capable of extracting information from it.

Moreover, the diversity of available data in terms of content, format and extension is huge. Indeed, the data available in microblogs such as Twitter are short and written without much concern for grammar, while review-related data are more extensive and follow stricter grammatical rules [1]. So it is also necessary to bear these differences in mind when attempting to perform any kind of analysis.

In this work, we will focus on two fields charged with dealing with the aforementioned problems, Opinion Mining (OM) and Information Fusion (IF). Opinion Mining (also known as *Sentiment Analysis* [2,3]) is a sub-field of text mining in which the main task is to extract opinions from content generated by Web users. Opinions play a fundamental role in the decision-making process of both individuals and organizations since they deeply influence people's attitudes and beliefs [4]. Such is the interest in harnessing the power to automatically detect and understand opinions that today this field is one of the most popular areas of research in the Natural Language Processing (NLP) and Computer Science communities, with more than 7000 articles published [5].

To mention some examples, mining opinions enables e-commerce businesses to gain deeper knowledge of their customers and products without having to pay for surveys [6], it allows politicians to understand the political sentiment of the community towards them without having to rely on polls [7], lets companies anticipate their stock trading volumes and financial returns [8], and helps strengthening the deliberation process in the public policy context [9].

Additionally, extracting opinions from reviews, blogs and microblogs, combined with the fusion of different sources of information presents several advantages such as higher authenticity, reduced ambiguity and greater availability [10]. Information Fusion is defined as "the study of efficient methods for automatically or semi-automatically transforming information from different sources and different points in time into a representation that provides effective support for human or automated decision making" [11]. Most of the research in Information Fusion has been done in fields related to the military where data is generated by electronic sensors, however there is growing interest in the fusion of data generated by humans (also called *soft* data) [10,12].

In this paper we attempt to review the state of the art in Opinion Mining studies that explicitly or implicitly use the fusion of information. Our aim is to provide both new and experienced researchers with insights on how to better perform the fusion

* Corresponding author. Tel.: +56 2 2978 4834; fax: +56 2 2689 7895.
E-mail addresses: jabalazs@ug.uchile.cl (J.A. Balazs), jvelasqu@dii.uchile.cl (J.D. Velásquez).

http://dx.doi.org/10.1016/j.inffus.2015.06.002
1566-2535/

process in an Opinion Mining context while also supplying enough information to help them understand both fields separately.

The remainder of this work is structured as follows: In Section 2 we show an overview of Opinion Mining by formally defining it, describing the usual process pipeline, explaining the different levels of analysis at which it performs, the different approaches that it uses and the most common challenges it faces. In Section 3 we review the state of the art in Opinion Mining combined with Information Fusion and present a simple framework for guiding the fusion process in the Opinion Mining context. Finally, in Section 4 we present some of the reviews that have been published both for Opinion Mining and Information Fusion.

2. Opinion Mining

Merriam-Webster's Online Dictionary[1] defines an opinion as a belief, judgement or way of thinking about something. Opinions are formed by the experiences lived by those who hold them. A consumer may look for another's opinion before buying a product or deciding to watch a movie, to gain insights into the potential experiences they would have depending on the decisions they make. Moreover, businesses could benefit from knowing the opinions of their customers by discovering cues on what aspects of a certain service to improve, which features of a determined product are the most valued, or which are new potential business opportunities [13,14]. In essence, a good Opinion Mining system could eliminate the need for polls and change the way traditional market research is done.

2.1. Definition

Opinion Mining is the field charged with the task of extracting opinions from unstructured text by combining techniques from NLP and Computer Science.

Liu [15] defines an opinion as a 5-tuple containing the target of the opinion (or *entity*), the attribute of the target at which the opinion is directed, the sentiment (or polarity) contained in the opinion which can be positive, negative or neutral, the opinion holder and the date when the opinion was emitted. Formally, an opinion is defined as a tuple:

$$(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$$

where e_i is the ith opinion target, a_{ij} is the jth attribute of e_i, h_k is the kth opinion holder, t_l is the time when the opinion was emitted and s_{ijkl} is the polarity of the opinion towards the attribute a_{ij} of entity e_i by the opinion holder h_k at time t_l.

Note that we described the sentiment contained in an opinion as positive, negative or neutral, notwithstanding it could also be numerically represented. For instance -5 could denote a very negative opinion while 5 a very positive one. Also, in case the analysis did not require much level of detail, the attributes of an entity could be omitted and denoted by *GENERAL* instead of a_{ij}.

Therefore the main objective of Opinion Mining is to find all the opinion tuples $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ within a document, collection of documents (called *corpus*) or across many corpora. Other works define Opinion Mining as "the task of identifying positive and negative opinions, emotions and evaluations" [16], "the task of finding the opinions of authors about specific entities" [5], "tracking the mood of the public about a particular product or topic" [17], or simply "the task of polarity classification" [18]. These definitions present different scopes and levels of granularity, however all of them can be adapted to fit Liu's opinion model.

[1] http://www.merriam-webster.com/dictionary/opinion (Visited May 11, 2015).

There are other approaches, like the one presented in [19], in which the authors attempt to classify emotional states such as "anger", "fear", "joy", or "interest" instead of just positive or negative. In this case, Liu's model could be enriched by adding another element to the opinion tuple model to represent this information.

2.2. Opinion Mining process: previous steps

The usual Opinion Mining process or pipeline usually consists of a series of defined steps [20–22]. These correspond to corpus or data acquisition, text preprocessing, Opinion Mining core process, aggregation and summarization of results, and visualization. In this paper we will give an overview of the first three. Particularly, in this section we will briefly review the two first steps previous to the core OM process: data acquisition and text preprocessing.

2.2.1. Data acquisition

The first step of any Opinion Mining pipeline is called corpus or data acquisition and consists of obtaining the corpus that is going to be mined for opinions. Currently there are two approaches to achieving this task. The first is through a website's Application Programming Interface (API) being Twitter's[2] one of the most popular [22–25]. The second corresponds to the use of Web crawlers in order to scrape the data from the desired websites [26–28]. Olston and Najork portray a robust survey of Web crawling in [29].

Both approaches present some advantages and disadvantages so there is a trade-off between using either. In [30] the authors briefly compare them.

With the API-based approach the implementation is easy, the data gathered is ordered and unlikely to change its structure, however it presents some limitations depending on the provider. For instance search queries to the Twitter REST API are limited to 180 per 15-min time window.[3] Additionally, the Streaming API has no explicit rate limits for downloading tweets, but is limited in other aspects such as the number of clients from the same IP address connected at the same time, and the rate at which clients are able to read data.[4] This approach is also subject to the availability of an API since not all websites provide one, and even if they do it might not present every needed functionality.

In contrast, crawler-based approaches are more difficult to implement, since the data obtained is noisier and its structure is prone to change, but have the advantage of being virtually unrestricted. Still, using these approaches requires to respect some good etiquette protocols such as the *robots exclusion standard*,[5] not issuing multiple overlapping requests to the same server and spacing these requests to prevent putting too much strain on it [29]. Furthermore, Web crawlers can prioritize the extraction of subjective and topically-relevant content. In [31], the authors propose a focused crawler that collects opinion-rich content regarding a particular topic and in [32] this work is further developed by proposing a formal definition for sentiment-based Web crawling along with a framework to facilitate the discovery of subjective content.

2.2.2. Text preprocessing

The second step in the OM pipeline is Text Preprocessing and is charged with common NLP tasks associated with lexical analysis [33]. Some of the most common techniques are:

Tokenization: task for separating the full text string into a list of separate words. This is simple to perform in space-delimited languages such as English, Spanish or French, but becomes

[2] https://dev.twitter.com/rest/public (Visited May 11, 2015).
[3] https://dev.twitter.com/rest/public/rate-limiting (Visited May 11, 2015).
[4] https://dev.twitter.com/streaming/overview/connecting (Visited May 11, 2015).
[5] http://www.robotstxt.org/robotstxt.html (Visited May 11, 2015).