# #HashtagWars: Learning a sense of humor (SemEval 2017)

Miha Pešič, Rok Zidarn

January 11, 2017

**Abstract**

Humor je eden izmed poglavitnih karakteristik človekovega obstoja in je morda celo nek pokazatelj inteligence. Čeprav ljudje povsem enostavno klasificiramo humor, je na področju računalništva in avtomatske detekcije ter klasifikacije humorja še vedno težava, kako oceniti humor. Trenutno je že precej težavno določiti binarni razred, bodisi je nekaj smešno bodisi ne, kaj šele podati oceno stopnje humorja, npr 10 - zelo smešno, 1 - povsem ne smešno. Težava izhaja predvsem iz tega saj je potrebno veliko predznanja, humor je tudi precej subjektiven, lahko se pojavi sarkazem ali ironija. Namen te naloge bi torej bil razviti nek klasifikator, ki bi čim bolje napovedal ali bo nekaj smešno ali ne. Izhajali bomo iz podatkov oddaje midnight, kjer imajo med humoristi tekmovanje imenovano #hashtagwars v katerem tekmujejo, kdo bo napisal čimbolj smešen tvit (ang. tweet) na določeno temo (ang. hashtag).

## 1 Introduction

Na področju umetne inteligence in procesiranja naravnega jezika je napovedovanje in ocenjevanje humorja v tekstu ena izmed bolj zanimivih nalog v zadnjem času. Ljudem je ponavadi povsem jasno, kdaj gre za šalo in kdaj ne, vendar sistemom kot so računalniki pa ne povsem. Potrebno je neko predznanje o sami temi šale, humor se lahko stopnjuje, nadaljuje, pojavi se lahko sarkazem, ironija. Včasih pride do pretiravanja ali igre besed (ang. puns). Kljub temu, da že obstajajo neki algoritmi in klasifikatorji trenutno lahko napovedujejo zgolj binarni razred, ali je tekst smešen (da/ne). Zaželjeno bi torej bilo, da bi lahko ocenili stopnjo humorja (1-10), vendar tukaj se bomo osredotočili na napovedovanje binarnega razreda, smešno (da/ne). Uporabili bomo podatkovno zbirko tvitov, spisanih v oddaji midnight, kjer so podatki o temi (hashtag-u), tekstu tvita in oceni. Ocena je določena z eno izmed treh vrednosti, 0 pomeni, da tvit ni smešen, 1 pomeni, da je tvit smešen, 2 pa, da je tvit najbolj smešen v podani temi. Seveda ni nujno, da bomo iz teh podatkov zgradili dovolj dober klasifikator, kajti to oceno je podala zgolj majhna skupina ljudi in potrebno je zopet pudariti, da je humor subjektivna zadeva.

### 1.1 Humor

Ljudje smo družabna bitja, za obstoj potrebujemo druge in humor je eden izmed načinov povezovanja, kakor tudi način sproščanja. Humor je tesno povezan s smehom oziroma z dobrim počutjem, kajti med tem procesom se sproščajo endorfini, ki jim drugače rečemo hormon sreče. Glede na to, da je humor subjektiven imamo ljudje različne okuse humorja, nekaterim je nekaj smešno, drugim ne. Kar vpliva na to je morda socialni status, vzgoja, družba ali kultura, zaradi česar ne moremo humorja enostavno opisati oziroma formalno zapisati.

### 1.2 Vrste pisanega humorja

Poznamo več vrst zapisanega humorja, oziroma več značilk, ki opisujejo vrsto humorja. Spodaj je zapisanih nekaj primerov:

1. Fonologija: “What do you use to talk to an elephant? An elly-phone."
2. Nasprotja: “Mythical Institute of Theology" - (MIT)
3. Negativna orientacija: “Money can’t buy your friends, but you do get a better class of enemy."
4. Stereotipi: “It was so cold last winter that I saw a lawyer with his hands in his own pockets."

5. Aliteracija: “Infants don’t enjoy infancy like adults do adultery."
6. Antonomija: “Always try to be modest and be proud of it!"

# 2 Methodology

## 2.1 Podatki

## 2.2 Predprocesiranje

## 2.3 Značilke

## 2.4 Klasifikacija

# 3 Results

# 4 Discussion